

(19) World Intellectual Property Organization
International Bureau



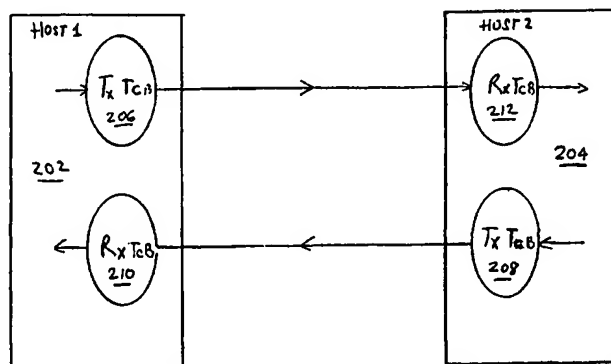
(43) International Publication Date
13 March 2003 (13.03.2003)

PCT

(10) International Publication Number
WO 03/021452 A1

- (51) International Patent Classification⁷: G06F 13/00 (74) Agent: NGUYEN, Joseph, A.; Ipsg, P.C., P.O. Box 61147, Sunnyvale, CA 94088 (US).
- (21) International Application Number: PCT/US02/27707
- (22) International Filing Date: 30 August 2002 (30.08.2002)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/316,651 31 August 2001 (31.08.2001) US
- (71) Applicant (for all designated States except US):
ADAPTEC, Inc. [US/US]; 691 South Milpitas Blvd., Milpitas, CA 95035 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): MUNNANGI, Sivakumar [IN/US]; 2266 Lenox Place, Santa Clara, CA 95054 (US); JAYAM, Ramkumar [IN/US]; 4573 Terra Place, San Jose, CA 95121 (US); KAPATKAR, Anil [IN/US]; 4719 Blue Ridge Dr., San Jose, CA 95129 (US); VENKATARAMAN, Srinivasan [IN/US]; 699 Patriot Place, Fremont, CA 94539 (US).
- (81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.
- (84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:
— with international search report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEMS AND METHODS FOR HIGH SPEED DATA TRANSMISSION USING TCP/IP



(57) Abstract: A method for exchanging data in a first host computer (202) coupled to a network. The network is also coupled to a second host computer (204). The method includes employing a first Tx Transmit Control Block (Tx TCB) (206) to facilitate transmitting first data associated with a bi-directional data flow from the first host computer (202) to the second host computer (204). The first data is transmitted using the TCP protocol. The first Tx TCB (206) is associated with the first host computer (202) and is configured for storing transmit-facilitating parameters associated with transmitting the first data using the TCP protocol. The method also includes employing a first Rx Transmit Control Block (Rx TCB) (210) to facilitate receiving second data associated with the bi-directional data flow. The second data is transmitted from the second host computer (204) to the first host computer (202) using the TCP protocol. The first Rx TCB (210) is associated with the first host computer (202) and is configured for storing receive-facilitating parameters associated with the receiving the second data using the TCP protocol. In accordance with this embodiment of the invention, the first Tx TCB (206) and the first Rx TCB (210) are employed to simultaneously service a transmit request pertaining to the first data and a receive request pertaining to the second data.

SYSTEMS AND METHODS FOR HIGH SPEED DATA TRANSMISSION
USING TCP/IP

BACKGROUND OF THE INVENTION

1 The present invention relates to data communication using a computer network. More particularly, the present invention relates to improved methods and apparatus for transmitting data among a plurality of computer systems.

2 The use of the transmission control protocol/internet protocol (TCP/IP) to facilitate the transmission of information between two or more computer systems via one or more networks is well known. When a given networked computer wishes to exchange information with another networked computer, a bi-directional data flow is established to allow information to be transmitted from one computer and received by the other computer. Generally speaking, the information is broken into packets to facilitate the transmission process. The TCP/IP protocol suite ensures that the information to be transferred is properly segmented and sent from the transmitting computer as packets, as well as properly received and assembled into the complete data file at the receiving computer.

3 As is well known, the transmission control protocol (TCP) corresponds to the transport layer (layer 4) of the OSI reference model. The transmission control protocol offers, among others, stream data transfer, multiplexing, full duplex operation, segmentation and reassembly, and efficient flow control. The internet protocol (IP) is a network layer (layer 3) protocol that provides, among others, addressing information and some control information that enables packets to be routed. The IP protocol has two primary responsibilities: providing connectionless, best-effort delivery of datagrams to a network, and providing fragmentation and reassembly of datagrams to support data links with different maximum transmission units (MTU) sizes. Together, these two protocols form the core of the internet protocol suite that enables reliable delivery of data via a network.

4 When two computers communicate via a computer network using the TCP/IP protocol, a data structure known as a transmission control block (TCB) is typically employed to facilitate data transmission, segmentation, reassembly, retransmission,

acknowledgement, and the like of datagrams in the bi-directional data flow between the communicating computers. The TCB is employed to track various parameters associated with the data transmit and receive process for a given data flow. Generally speaking, there is one transmission control block per data flow at each host computer system (i.e., the computer system involved in the communication at each end of a data flow). Each TCB is uniquely identified by its TCP source port, TCP destination port, IP source address, and/or IP destination address.

5 In the prior art, a transmission control block in a host computer is employed, for a given data flow, to facilitate both the transmission of data from that host computer and the receiving of data into that host computer. Fig. 1 illustrates this situation wherein a TCB 102 is employed to facilitate both the transmission and the receiving of data for host computer 104. Likewise, a TCB 106 is employed to facilitate both the transmission and the receiving of data for host computer 108. However, if a transmission control block is busy servicing a transmission request in a host computer, it is unavailable for use to facilitate the receiving of data in that host computer until the transmit task is finished. Accordingly, a bottleneck exists which limits the transmission bandwidth between the two host computers.

6 If the data transmission speed between host computer 104 and host computer 108 is relatively low compared to the speed at which data is processed within the host computers, this bottleneck may be tolerable. As the transmission bandwidth between host computers increase, this bandwidth bottleneck increasingly becomes a critical issue. As bandwidth approaches 1Gbits/sec, 10Gbits/sec, or even higher for enterprise networking, and up to 40Gbits/sec or higher among network routers, it is clear that the bandwidth bottleneck needs to be resolved if TCP/IP is to remain a viable protocol suite for networking going forward.

7 In view of the foregoing, there is desired improved methods and apparatus for relieving the bandwidth bottleneck associated with the prior art transmission control block and for improving the data transmission speeds when two or more computers communicate using the TCP/IP protocol via a computer network.

SUMMARY OF THE INVENTION

8 The invention relates, in one embodiment, to a method for exchanging data in a first host computer coupled to a network. The network is also coupled to a second host computer. The method includes employing a first Tx Transmit Control Block (Tx TCB) to facilitate transmitting first data associated with a bi-directional data flow from the first host computer to the second host computer. The first data is transmitted using the TCP protocol. The first Tx TCB is associated with the first host computer and is configured for storing transmit-facilitating parameters associated with the transmitting the first data using the TCP protocol. The method also includes employing a first Rx Transmit Control Block (Rx TCB) to facilitate receiving second data associated with the bi-directional data flow. The second data is transmitted from the second host computer to the first host computer using the TCP protocol. The first Rx TCB is associated with the first host computer and is configured for storing receive-facilitating parameters associated with the receiving the second data using the TCP protocol. In accordance with this embodiment of the invention, the first Tx TCB and the first Rx TCB are employed to simultaneously service a transmit request pertaining the first data and a receive request pertaining the second data.

9 In another embodiment, the invention relates to circuitries for facilitating data exchange via a network. The circuitries are associated with a first host computer coupled to said network. The network is also coupled to a second host computer. The circuitries include means for facilitating transmitting first data associated with a bi-directional data flow from the first host computer to the second host computer. The first data is transmitted using the TCP protocol. The means for facilitating transmitting is associated with the first host computer and is configured for storing transmit-facilitating parameters associated with transmitting the first data using the TCP protocol. The circuitries also include means for facilitating receiving second data associated with the bi-directional data flow. The second data is transmitted from the second host computer to the first host computer using the TCP protocol. The means for facilitating receiving is associated with the first host computer and is

configured for storing receive-facilitating parameters associated with receiving the second data using the TCP protocol. In accordance with this embodiment of the invention, the means for facilitating transmitting and the means for facilitating receiving are employed simultaneously to facilitate simultaneous transmitting the first data and receiving the second data by the circuitries.

- 10 In yet another embodiment, the invention relates to a method on a network interface card (NIC card) coupled to a first host computer for exchanging data via a network coupled to the first host computer and a second host computer. The method includes transmitting, using transmit-facilitating parameters stored in a first Tx Transmit Control Block (Tx TcB) associated with the NIC card, first data associated with a bi-directional data flow. The first data is transmitted from the first host computer to the second host computer. The first data is transmitted using the TCP protocol. The method also includes receiving, using receive-facilitating parameters stored in a first Rx Transmit Control Block (Rx TcB) associated with the NIC card, second data associated with the bi-directional data flow. The second data is transmitted from the second host computer to the first host computer. The second data is transmitted using the TCP protocol. In accordance with this embodiment of the invention, the Tx TcB and the Rx TcB are employed to facilitate simultaneous transmitting the first data and receiving the second data by the NIC card.

- 11 These and other features of the present invention will be described in more detail below in the detailed description of the invention and in conjunction with the following figures.

BRIEF DESCRIPTION OF THE DRAWINGS

12 The present invention is illustrated by way of example, and not by way
of limitation, in the figures of the accompanying drawings and in which like reference
numerals refer to similar elements and in which:

13 Fig. 1 illustrates this situation wherein a prior art TCB is employed to facilitate
both the transmitting and the receiving processes for host computer, thereby enabling
only one process to occur at any given time.

14 Fig. 2 shows, in accordance with one embodiment of the present invention, a
simplified diagram showing a first host computer exchanging data with a second host
computer using the inventive transmit control block (Tx TCB) and the inventive
receive control block (Rx TCB) of the present invention.

15 Fig. 3 shows, in accordance with one embodiment of the present invention, a
transmit control block data structure for facilitating transmitting data.

16 Fig. 4 illustrates, in accordance with one embodiment of the present invention,
a transmit window, which conceptually represents the amount of data allowed to be
transmitted from the transmitting host computer for a given data flow at any given
time.

17 Fig. 5 illustrates, in accordance with one aspect of the present invention, a
retransmit queue and a transmit pending queue.

18 Fig. 6 is a block diagram illustrating, in accordance with one embodiment of
the present invention, the transmit operations involving the transmit control block of
the present invention.

19 Fig. 7 is a simplified flowchart illustrating, in accordance with one
embodiment of the present invention, how the transmit network protocol processor
may employ the transmit control block to service a request to transmit data from the
host application program associated with the transmitting host computer.

- 20 Fig. 8 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how the transmit network protocol processor may employ the transmit control block to service a request to retransmit data from retransmit queue.
- 21 Fig. 9 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how the transmit network protocol processor may employ the transmit control block to send an acknowledgement to the other transmitting host computer.
- 22 Fig. 10 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how the transmit network protocol processor may involve the transmit control block in updating the transmit window and calculating performance data responsive to a received acknowledgement message sent from the other transmitting host computer.
- 23 Fig. 11 illustrates, in accordance with one embodiment of the present invention, an exemplary receive control block (Rx TCB) data structure.
- 24 Figs. 12A and 12B show, in accordance with embodiments of the present invention, exemplary steps for receiving data packets using the receive control block (Rx TCB).
- 25 Fig. 13 illustrates, in accordance with one embodiment of the present invention, the steps taken by the receive process in response to the receipt of an acknowledgement packet.
- 26 Fig. 14 illustrates, in accordance with one embodiment of the present invention, the steps taken by the receive process in response to the receipt of one or more data packets that contain data other than acknowledgement for data previously sent.
- 27 Fig. 15 illustrates, in accordance with one embodiment of the present invention, a receive transmit control block (Rx TCB) data structure for facilitating receiving data at the target device.

- 28 Fig. 16 illustrates, in accordance with one embodiment of the present invention, a simplified diagram showing a Rx TCB whose out-of-order packets are partially reordered.
- 29 Fig. 17 illustrates, in accordance with another embodiment of the present invention, a simplified diagram showing an exemplary Rx TCB that includes the partial reordering feature for out-of-order packets.
- 30 Fig. 18 illustrates, in accordance with another aspect of the present invention, a simplified diagram illustrating the look-ahead feature with regard to the partial reordering of out-of-order packets in a Rx TCB.
- 31 Appendix A, Fig. A1 is a block diagram illustrating the elements of a prior art TCP data segment;
- 32 Appendix A, Fig. A2a illustrates the layout in memory of a TCB data structure according to a preferred embodiment of the present invention;
- 33 Appendix A, Fig. A2b is a flow chart illustrating Reorder Buffer (ROB) update and processing;
- 34 Appendix A, Figs. A3a-A3g list the preferred transmit variables, receive variables, and re-order variables used in a plurality of algorithms needed to implement the TCP protocol;
- 35 Appendix A, Fig. A4 is a flow diagram illustrating the steps for transmitting a stream of data, comprising a plurality of data segments, from a sending computer to a receiving computer according to a preferred embodiment of the present invention;
- 36 Appendix A, Fig. A5 is a block diagram illustrating an ASIC having microcode driven processing engines for TCP/IP processing according to the present invention;
- 37 Appendix A, Fig. A6 illustrates how Fibre Channel frame to Gigabit Ethernet frame processing is performed in the ASIC of Appendix A, Fig. A5; and
 Fig. A7 illustrates how Gigabit Ethernet frame to Fibre Channel frame processing is performed in the ASIC of Fig. A5.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

38 The present invention will now be described in detail with reference to a few preferred embodiments thereof as illustrated in the accompanying drawings. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art, that the present invention may be practiced without some or all of these specific details. In other instances, well known process steps and/or structures have not been described in detail in order to not unnecessarily obscure the present invention.

39 Fig. 2 shows, in accordance with one aspect of the present invention, a simplified diagram showing a host computer 202 exchanging data with a host computer 204 using the inventive transmit control block (Tx TCB) and the inventive receive control block (Rx TCB) of the present invention. As shown in Fig. 2, there is one transmit control block (Tx TCB) for each data flow in each of the host computers. The transmit control block (Tx TCB) is provided in a host computer and is employed during the data transmit process. Thus, each bi-directional flow of data is associated with two transmit control blocks (Tx TCBs) if both host computers at the two ends of the data flow are involved in transmitting data to one another. This is shown by transmit control blocks 206 and 208 in host computers 202 and 204 respectively. For receiving data, each host computer at the two ends of the bi-directional data flow also has a receive control block (Rx TCB). Thus, each bi-directional flow of data is associated with two receive control blocks (Rx TCBs) if both host computers at the two ends of the data flow are permitted to receive data from one another. This is shown by receive control blocks 210 and 212 in host computers 202 and 204 respectively.

40 Within each host computer, since each bi-directional data flow is provided with a separate data control block for the transmit process (i.e., the transmit control block or Tx TCB) and a separate data control block for the receive process (i.e., the receive control block or Rx TCB), that host computer can simultaneously process

transmit and receive requests using two different data control blocks. For example, host computer 202 can service simultaneously a transmit request using Tx TCB 206 and a receive request using Rx TCB 210. As the term is employed herein, processes occurring simultaneously denotes that the sub-step or sub-process of one can be executed even before the other process is finished. For example, two processes executed in two different processing circuits (such as two parallel processors) can occur simultaneously. Further, two processes executed in two different threads by a single processing circuit can be deemed to occur simultaneously even if the processing circuit, by its nature, can execute only one instruction at a time.

41 In the context of the present invention, the transmit request pertaining to certain packets pertaining to a bi-directional data flow can be serviced simultaneously with the servicing of a receive request pertaining to other packets of that bi-directional data flow. This is different from the situation in Fig. 1 wherein the host computer, such as host computer 104, must wait until the servicing of the transmit request is done and the prior art TCB (such as TCB 102) is released before having access to that TCB in order to service the receive request (or vice versa).

42 As the term is employed herein, a host computer refers to a computer-implemented device having a processor and at least one I/O port for transmitting and receiving data (hence the term "host"). It is not necessary, as the term is employed herein, for a host computer to have a keyboard or other user data entry arrangements. Further, it is not necessary, as the term is employed herein, for a host computer to have a data display device, such as a computer display screen. As such, a host computer, as the term is employed herein, may include a storage area network (SAN) arrangement, network attached storage (NAS), a data storage device having a processor for communicating with other devices via a network, and/or indeed any other computer-implemented device having a processor and capable of communicating via a network.

43 Fig. 3 shows, in accordance with one embodiment of the present invention, a transmit control block (Tx TCB) 300 data structure for facilitating transmitting data. Referring now to Fig. 3, transmit control block (Tx TCB) 300 has, among others, four major groups of transmit-facilitating parameters (i.e., parameters employed to

facilitate the transmit process): TX header, TX window management, TX queues management, and TX timer management. The more important TX header parameters include TCP source port (TCP SP) 302, TCP destination port (TCP DP) 304, IP source address (IP SA) 366, IP destination address (IP DA) 306, the sequence number to be sent next (SND_NXT) 322, the window size of the transmitting host computer (SND_WIN) 316, the Optimal Maximum Segment Size (SEG_OPT_MSS) 318, the current congestion window (ICNG_WIN/CNG_WIN) 324.

44 The more important TX window management parameters include the current congestion window (ICNG_WIN/CNG_WIN) 324, the old unacknowledged sequence number (OLD_UNA) 320, the sequence number to be sent next (SND_NXT) 322, the maximum sequence number to be sent (SND_MAX) 326, the window size of the transmitting host computer (SND_WIN) 316.

45 The more important TX queue management parameters include Transmit Pending Read Pointer (TXP_RD_PTR) 346, Transmit Pending Write Pointer (TXP_WR_PTR) 344, Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350, Retransmit Read Pointer (TXP_RTX_PTR) 348, Transmit Queue Base Address (TXP_BASE_ADDR) 330, and Transmit Queue Maximum Address (MAX_TXP_ADDR) 342.

46 The more important TX timer management parameters include the Retransmission Time Out value RTO 334, the Frame Transmit Timer FST 340, the Smoothened Round Trip Time (SRTT) 338, the Measured Round Trip Time (MRTT) 336, and the Sequence number of the segment for which timer measurements are kicked off (TMR_SEQ_NUM) 332.

47 The operation of transmit control block (Tx TCB) 300 may be better understood with reference to the figures that follow. Integral to the concept of a transmit control block (Tx TCB) are the concepts of transmit window management and queue management. Fig. 4 illustrates in a simplified format a transmit window 402, which conceptually represents the amount of data allowed to be transmitted from the transmitting host computer for a given data flow at any given time. Transmit window 402 may have a size of, for example, 64Kb although the exact value may vary

with implementations. Whenever there is a request to transmit a packet (generally from the host application software), the packet is buffered and transmit window 402 is checked (via parameter SND_WIN 316 in the transmit control block 300) to see whether there is sufficient transmit bandwidth to send out the packet.

48 Transmit window 402 may have a slow start feature; that is, transmit window 402 starts out being rather small, e.g., up to first window size 404 (generally 1 MSS or 1 Maximum Segment Size). This allows a small amount of data to be transmitted from the host computer in the beginning. If the host computer at the other end sends an acknowledgement timely, transmit window 402 gradually opens up further to allow more data to be transmitted through, up to a maximum size indicated by maximum window size 406. If an expected acknowledgement is not timely received, transmit window 402 shrinks to its minimum size again, or to some value smaller than the value it had when the acknowledgement period expires. In this manner, transmit window 402 performs the function of congestion control at the transmitting host computer.

49 Fig. 5 illustrates, in accordance with one aspect of the present invention, a retransmit queue 502 and a transmit pending queue 504. Transmit pending queue 504 is employed to queue transmit elements, which may include either the packet to be transmitted from the host computer associated with the transmit control block (Tx TCB) or, more commonly, a pointer to that packet. A transmit element may also include data associated with the packet to be transmitted such as the starting sequence number and the length of the packet.

50 If the transmit window size is sufficient to transmit a packet, that packet is transmitted and a copy (or its pointer and/or associated data) is placed into the retransmit queue 502. On the other than, a packet (or its pointer and associated data) to be transmitted is placed into transmit pending queue 504 when the transmit window size is insufficient to transmit that packet. Transmit pending queue 504 is managed by two pointers: Transmit Pending Read Pointer (TXP_RD_PTR) 346 and Transmit Pending Write Pointer (TXP_WR_PTR) 344. Transmit Pending Read Pointer (TXP_RD_PTR) 346 points to the top of the transmit queue and specifically to the queue element to be sent next when the transmit window is available. Transmit

Pending Write Pointer (TXP_WR_PTR) 344 is the tail of the transmit queue and represents the last inserted transmit element. When a packet is sent out from transmit pending queue 504, Transmit Pending Read Pointer (TXP_RD_PTR) 346 moves toward Transmit Pending Write Pointer (TXP_WR_PTR) 344. Thus, the difference between Transmit Pending Write Pointer (TXP_WR_PTR) 344 and Transmit Pending Read Pointer (TXP_RD_PTR) 346 represents the size of the transmit queue at any given point in time, which reflects the number of pending packets to be sent.

51 After being sent, a packet is moved to retransmit queue 502. In retransmit queue 502, the packet awaits acknowledgement from the other host computer before being dequeued. Retransmit queue 502 is managed by two pointers: Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 and Retransmit Read Pointer (TXP_RTX_PTR) 348.

52 In most cases, Retransmit Read Pointer (TXP_RTX_PTR) 348 and Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 point at the same retransmit queue element. During retransmit, Retransmit Read Pointer (TXP_RTX_PTR) 348 advances while Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 stays, acting as a shadow pointer. Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 is advanced only when an acknowledgement is received for the point retransmitted. If all retransmitted packets are acknowledged, Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 and Retransmit Read Pointer (TXP_RTX_PTR) 348 will catch up with one another and will point at the same queue element again.

53 For convenience, retransmit queue 502 and transmit pending queue 504 may be implemented together as a circular queue and occupy a contiguous block of memory. However, the use of Retransmit Read Pointer (TXP_RTX_PTR) 348, Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350, Transmit Pending Read Pointer (TXP_RD_PTR) 346, and Transmit Pending Write Pointer (TXP_WR_PTR) 344 allow this retransmit queue 502 and transmit pending queue 504 to be managed as two different logical queues.

54 Fig. 6 is a block diagram illustrating, in accordance with one embodiment of the present invention, the transmit operations involving the transmit control block (Tx

TCB) of the present invention. Referring to Fig. 6, there is shown a transmit network protocol processor (TX NPP) 602, representing the processor servicing the transmit operations. In the example of Fig. 6, transmit network protocol processor (TX NPP) 602 represents a micro-code driven engine, or a plurality of micro-code driven engines, which takes in, among others, four different types of requests: Request to send data, request to retransmit, request to send acknowledgement (ACK), and request to update transmit window based on acknowledgement received. In general, these four requests may be serviced concurrently by transmit network protocol processor (TX NPP) 602 using four independently executing threads.

55 The Requests to send data are typically generated by the application software associated with the transmitting host computer. These requests are queued up in a queue 604 while awaiting to be serviced by transmit network protocol processor (TX NPP) 602. In one embodiment, the content of each queue element of queue 604 includes the pointer to the transmit control block (Tx TCB) associated with the data flow, the pointer to the buffer element where the data is stored, and the length of the data to be transmitted.

56 Since there may be multiple data flows between a given host computer and another host computer or other host computers, each host computers may have multiple transmit control blocks (Tx TCBs), each associated with a data flow. The plurality of transmit control blocks (Tx TCBs) are shown in Fig. 6 by reference number 612. The transmit control block (Tx TCB) for a given data flow may be found using, for example, the TCP source port (TCP SP), the TCP destination port (TCP DP), the IP destination address (IP DA), and/or the IP source address (IP SA).

57 The requests to retransmit are typically generated when a retransmit timer expires after some predefined time has passed since a given packet was transmitted, and an acknowledgement has not yet been received. Since there may be multiple data flows between a given host computer and another host computer or other host computers, each host computers may have multiple retransmit timers. The plurality of retransmit timers are shown in Fig. 6 by reference number 614. These requests are queued up in a queue 606 while awaiting to be serviced by transmit network protocol processor (TX NPP) 602. In one embodiment, the content of each queue element of

queue 606 includes the pointer to the transmit control block (Tx TCB) associated with the data flow.

58 The requests to send acknowledgement (ACK) are generated responsive to the successful receipt of data transmitted from the other host computer. These requests are queued up in a queue 608 while awaiting to be serviced by transmit network protocol processor (TX NPP) 602. In one embodiment, the content of each queue element of queue 608 includes the pointer to the transmit control block (Tx TCB) associated with the data flow, the acknowledgement number associated with the data received, and the window size associated with the data received.

59 The requests request to update transmit window based on acknowledgement received are generated responsive to the acknowledgement received. These requests are queued up in a queue 610 while awaiting to be serviced by transmit network protocol processor (TX NPP) 602. In one embodiment, the content of each queue element of queue 610 includes the receive window size, the acknowledge number received, the pointer to the Tx TCB, and optionally time stamp and SACK information.

60 In one embodiment, these queues 604, 606, 608, and 610 are serviced by transmit network protocol processor (TX NPP) 602 in a round-robin manner although any other scheme to ensure that the requests in these queues are serviced in an efficient manner may also be employed.

61 The operation of the transmit network protocol processor (TX NPP) 602 and the role of the transmit control block (Tx TCB) in servicing the requests associated with the transmitting operations may be better understood with reference to the flowcharts of Figs. 7, 8, 9, and 10 herein. In one embodiment, the processes illustrated in Figs. 7-10 represent processes offloaded from the host processor of the host computer in order to improve data throughput. By way of example, these processes may be implemented by circuitry comprising one or more network processors and/or embedded processors and/or co-processors operating in parallel with the host processor of the host computer. The aforementioned circuitry may be implemented in an integrated manner with the host computer or may be implemented

on a plug-in card, such as a network interface card (NIC), that is configured to be detachably coupled to the bus of the host processor.

62 Fig. 7 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how transmit network protocol processor (TX NPP) 602 may employ the transmit control block (Tx TCB) to service a request to transmit data from the host application program associated with the transmitting host computer. Once the application requests to send out data, that data is fetched and placed into buffer memory. Each request is associated with a particular data flow, and hence a particular transmit control block (Tx TCB). Using the TCP source port (TCP SP), the TCP destination port (TCP DP), the IP destination address (IP DA) and/or the IP source address (IP SA), the transmit control block (Tx TCB) associated with the request is then obtained. The request is then placed into queue 604 to be serviced by transmit network protocol processor (TX NPP) 602 as discussed earlier.

63 In block 702, the method first check to see whether the transmit window is available to send the requested data. With reference to Fig. 3, the current window size is obtained from the transmit control block (Tx TCB), and more particularly, from the parameter SND_WIN 316 of the associated transmit control block (Tx TCB). If the window is available to transmit the requested data packet, the method performs certain housekeeping functions prior to sending the requested data packets onward to be sent out. Thus, the method proceeds to block 704 to update the transmit window size of the transmitting host computer (SND_WIN) 316, i.e., to reduce the size of the transmit window as part of transmit window management. Further, the method may also update the sequence number to be sent next (SND_NXT) 322 so that that the sequence number to be sent next (SND_NXT) 322 can reflect the next sequence number to be used for transmitting the next packet. Additionally, the method may also update TMR_SEQ_NUM 332, thereby starting a timer so that the delay can be measured once the acknowledgement is received.

64 In block 706, the method also places the transmitted packet(s) into retransmit queue 502, where they will stay until an acknowledgement is received. Thus, the Transmit Pending Read Pointer (TXP_RD_PTR) 346 is advanced. Furthermore Frame Transmit Timer FST 340 is also updated.

65 In block 708, the retransmit timer 614 is started for this data flow if it has not been started earlier. In block 710, the request, along with the transmit control block (Tx TCB) pointer, is send from transmit network protocol processor (TX NPP) 602 to Header Preparation Network Protocol Processor (HDR NPP) 616 (see Fig. 6) to prepare the header along with the data to send to the receiving host computer.

66 In one embodiment, the following information is transmitted from transmit network protocol processor (TX NPP) 602 to header preparation processor (HDR NPP) 616 to facilitate the preparation of a data packet (including the packet header) for transmission: pointer to the buffer where the data resides, transmit control block (Tx TCB) pointer, sequence #, length, acknowledge #, window size, and flags such as PSH, ACK, or URG (which are used in the TCP header). Once transmit network protocol processor (TX NPP) 602 sends this information to header preparation processor (HDR NPP) 616, the process for servicing the current transmission request ends.

67 On the other hand, if the transmit window is not available (as determined in block 702), the method proceeds from block 702 to block 714 to queue the packet on the transmit pending queue. Thus Transmit Pending Write Pointer (TXP_WR_PTR) 344 is updated to reflect the fact that the packet has been queued at the tail of the pending transmit queue.

68 Fig. 8 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how transmit network protocol processor (TX NPP) 602 may employ the transmit control block (Tx TCB) to service a request to retransmit data from retransmit queue 502. The process of Fig. 8 is invoked when the retransmit timer RTX TMR 614 associated with this data flow has expired. In this case, the process will send out one or more MSS-size packets (maximum segment size) from the retransmit queue, the exact number of MSS to be sent depends on ICNG_WND (324), which tracks congestion.

69 In block 804, the packets to be retransmitted (the exact number of which is determined in block 802) are sent out. As far as the management of retransmit queue 502 is concerned, this involves keeping Retransmit Shadow Pointer

(TXP_SHD_RD_PTR) 350 at the head of retransmit queue 502 and moving Retransmit Read Pointer (TXP_RTX_PTR) 348 to reflect the number of packets sent out.

70 In block 806, a set of data is acquired from both the retransmit queue element(s) associated with the packet(s) to be retransmitted and the associated transmit control block (Tx TCB). For example, data such as the buffer pointer, the sequence number, the length of the data to be retransmitted may be obtained from the retransmit queue element. From the associated transmit control block (Tx TCB), the data obtained may include the acknowledgement number, the receive window size (354 of Fig. 3). Furthermore, since a retransmit is about to be performed, the process adjusts the current congestion window (ICNG_WIN/CNG_WIN) 324. As mentioned earlier, the transmit window is reduced after data is transmitted or retransmitted until acknowledgement is received.

71 In block 808, the retransmit request along with the transmit control block (Tx TCB) pointer is send from transmit network protocol processor (TX NPP) 602 to header preparation processor (HDR NPP) 616 (see Fig. 6) to prepare the header along with the retransmit data to send to the receiving host computer. In one embodiment, the following information is transmitted from transmit network protocol processor (TX NPP) 602 to header preparation processor (HDR NPP) 616 to facilitate the preparation of a data packet (including the packet header) for transmission: pointer to the buffer where the data resides, transmit control block (Tx TCB) pointer, sequence #, length, acknowledge #, window size, and TCP-related flags such as PSH, ACK, or URG. Once transmit network protocol processor (TX NPP) 602 sends this information to header preparation processor (HDR NPP) 616, the process for servicing the current retransmit request of Fig. 8 ends.

72 Fig. 9 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how transmit network protocol processor (TX NPP) 602 may employ the transmit control block (Tx TCB) to send an acknowledgement to the other transmitting host computer. The process of Fig. 8 is invoked after the host computer has successfully received data packet(s) from the other transmitting host computer. In block 902, it is recognized that one or more data

packet(s) has been successfully received at the host computer. Based on the received data sequence, the Rx process prepares the ACK # to be sent and the receive window size and queues the request to TX NPP to send the ACK. Furthermore, the information in the received data packet(s) is also employed to obtain a pointer to the transmit control block (Tx TCB) associated with this data flow. In block 904, the current sequence number is obtained from the sequence number to be sent next (SND_NXT) 322.

73 In block 906, the information obtained in blocks 902 and 904 is sent to header preparation processor (HDR NPP) 616 to prepare a special acknowledgement packet to be sent to the other transmitting host computer. In one embodiment, the following information is transmitted from transmit network protocol processor (TX NPP) 602 to header preparation processor (HDR NPP) 616 to facilitate the preparation of an acknowledgement packet (including the packet header) for transmission: transmit control block (Tx TCB) pointer, sequence #, length, acknowledge #, window size. Once transmit network protocol processor (TX NPP) 602 sends this information to header preparation processor (HDR NPP) 616, the process for sending an acknowledgement of Fig. 9 ends.

74 Fig. 10 is a simplified flowchart illustrating, in accordance with one embodiment of the present invention, how transmit network protocol processor (TX NPP) 602 may involve the transmit control block (Tx TCB) in updating the transmit window and calculating performance data responsive to a received acknowledgement message sent from the other transmitting host computer. Generally speaking, there are two main tasks that need to be handled: (1) updating the parameters that manage the transmit window and, as a result, sending out data packets if the updated window size allows such transmitting, and (2) calculating performance data based on the information in the transmit control block (Tx TCB) and the information received in the acknowledgement message.

75 In block 1002, the TX window parameters are updated. As can be seen, the parameters updated are the window size of the transmitting host computer (SND_WIN) 316, the old unacknowledged sequence number (OLD_UNA) 320, the Frame Transmit Timer FST 340, and TMR_SEQ_NUM 332. In block 1004,

performance data is calculated and the following parameters are updated:
Retransmission Time Out value RTO 334, Smoothened Round Trip Time (SRTT)
338, Measured Round Trip Time (MRTT) 336.

76 In block 1006, the process releases any data packet(s) held for retransmit (as a result of an earlier transmit from either transmit pending queue 504 or retransmit queue 502) if such data packet(s) pertain to the received acknowledgement message. Thus, the process updates Retransmit Shadow Pointer (TXP_SHD_RD_PTR) 350 and/or Retransmit Read Pointer (TXP_RTX_PTR) 348 as appropriate. In block 1008, the process determines whether there is any pending data packets in transmit pending queue 504 to be sent out.

77 In one embodiment, if the difference between the head of transmit pending queue 504 (i.e., Transmit Pending Read Pointer (TXP_RD_PTR) 346) and the tail of transmit pending queue 504 (i.e., Transmit Pending Write Pointer (TXP_WR_PTR) 344) is greater than zero, the process deems there is pending data packets awaiting transmittal. In this case, the process outputs a request to cause the transmit network protocol processor (TX NPP) to invoke the process of Fig. 7 to perform data transmittal. This is seen in block 1010. On the other hand, if there is no more pending data packets in transmit pending queue 504 (as determined in block 1008), the process of Fig. 10 ends at block 1012.

78 As mentioned earlier, the use of a separate transmit control block (Tx TCB) and a separate receive control block (Rx TCB) to handle the transmitting and receiving tasks simultaneously at a given host computer for a given data flow eliminates the bandwidth bottleneck found in the prior art when a transmit control block is employed for both the transmitting and receiving tasks for the data flow. Fig. 11 illustrates, in accordance with one embodiment of the present invention, an exemplary receive control block (Rx TCB) data structure 1100. Rx TCB 1100 is employed to store receive-facilitating parameters, i.e., parameters employed to facilitate the receive process. Receive control block (Rx TCB) 1100 has, among others, parameters to enable a plurality of functions, including: reordering out-of-order packets, allowing the receive network protocol processor (RCV NPP), instead of the

transmit network protocol processor (TX NPP), to send out an acknowledgement packet, and maintaining Rx window size on a connection-by-connection basis.

79 With respect to out-of-order packet reordering, there is provided, in accordance with one aspect of the present invention, a method for partially reordering out-of-order packets so as to render the packet receiving process more bandwidth efficient. Suppose the receive network protocol processor (RCV NPP) was expecting packets having sequence numbers 1 to 1000. However, before the packets having sequence numbers 1-1000 are received, the receive network protocol processor (RCV NPP) receives instead packets having sequence numbers 1002 to 64,000. In some prior art implementations, these out-of-order packets would be discarded, requiring the transmitting host computer to send them again after the expected packets (i.e., those having sequence numbers 1-1000) are received.

80 In accordance with one aspect of the present invention, the receive network protocol processor (RCV NPP) would keep the packets having sequence numbers 1002 to 64,000, partially reordering them as necessary, but store them in the out-of-order buffers instead of sending them to the host software within the receiving host computer. The partially ordered packets may be stored, for example, in a linked list using the sequence number of the packets as the key.

81 By partially reordering the out-of-order packets (the re-ordering is only partial since packets having sequence numbers 1-1000 are still missing), the packets having sequence numbers 1-64,000 can be very quickly assembled once the packets having sequence numbers 1-1000 arrive. This rapid assembly permits the receive network protocol processor (RCV NPP) to more quickly send the complete data to the higher layer software. Furthermore, by reducing the amount of retransmission that the transmitting host computer has to perform, the invention effectively increases the bandwidth of the communication channel between the two host computers. This aspect of the invention is discussed in greater detail in a co-pending application entitled "Methods And Apparatus For Handling Out-Of-Order Packets," filed on even date and incorporated by reference (Attorney Docket No. ATECP002-R4/SNG-028A).

82 In receive control block (Rx TCB) 1100, there is shown a plurality of reorder buffers (ROBs): ROB1, ROB2, ROB3, and ROB4. There may be as many reorder buffers as desired although only four are shown. Each ROB is employed to store parameters associated with an out-of-order packet. The parameters associated with each ROB includes the following: Pointer to the data (PKT_PTR), Storage Transport Layer Header (STL, which is employed for Fiber Channel Frame Reassembly), Length of current packet (PKT_LEN), and Sequence number of current packet (SEQ_NUM).

83 With reference to ROB1, for example, these four parameters are shown by reference numbers 1104, 1106, 1108, and 1110 respectively. To provide for storage flexibility and to avoid making each receive control block (Rx TCB) unduly large, the out-of-order buffers may be extended into the memory space of the host computer system. Thus, if a greater number of reorder buffers are required beyond what is provided in the receive control block (Rx TCB) data structure, a pointer (ROB EXTENSION PTR) 1112 is provided, which points to a location in the memory space of the host computer system where the extension reorder buffers may be found.

84 The role of the receive control block (Rx TCB) in the data receiving operations may be better understood with reference to Figs. 12A-14 that follow. In one embodiment, the processes illustrated in Figs. 12A-14 represent processes offloaded from the host processor of the host computer in order to improve data throughput. By way of example, these processes may be implemented by circuitry comprising one or more network processors and/or embedded processors and/or co-processors operating in parallel with the host processor of the host computer. The aforementioned circuitry may be implemented in an integrated manner with the host computer or may be implemented on a plug-in card, such as the aforementioned network interface card (NIC), that is configured to be detachably coupled to the bus of the host processor.

85 Fig. 12A shows, in accordance with one embodiment of the present invention, exemplary steps for receiving data packets using the receive control block (Rx TCB). In block 1202, the sequence number(s) of the received data packet(s) are compared with the expected sequence number(s), which are stored by RCV_NXT 1120 in the receive control block (Rx TCB). The receive control block (Rx TCB) for the received packet may be found using information such the TCP source port, the TCP destination

port, the IP source address, and/or the IP destination address. If the two match, then the received packet(s) are deemed in order and the method proceeds to block 1204. On the other hand, if there is a discrepancy, the received packet(s) are deemed out of order, and the method proceeds to block 1248, which is described further in Fig. 12B herein.

86 In block 1204, the method updates RCV_NXT 1120, which updates the sequence number expected for the next packet(s) received. Furthermore, the receive window is updated by updating RCV_WIN 1122, which reduces the receive window size temporarily until data is sent from the receive network protocol processor (RCV NPP) to the application software in the receiving host computer.

87 Once updating is performed in block 1204, the method proceeds to block 1206 wherein the method decides whether the received packet pertains to received acknowledgement for data sent earlier from this host computer, or whether the received data packet pertains to data, other than an acknowledgement, sent from the other transmitting host computer. If the received data packet pertains to an acknowledgement associated with data sent earlier from this host computer, the method proceeds to block 1208 wherein the received acknowledgement is sent to the transmit network protocol processor (TX NPP) so that the transmit network protocol processor (TX NPP) can update its window, calculate performance data, and the like. This aspect has been discussed earlier in connection with Fig. 10 herein.

88 On the other hand, if the received data packet pertains to data, other than an acknowledgement, sent from the other transmitting host computer, the method proceeds to block 1210 wherein the data is moved into the host computer's memory, e.g., into the host buffer memory space, for use by the host software application. In blocks 1212 and 1214, the acknowledgement procedure is undertaken. In one embodiment, a cumulative acknowledgement scheme is employed (but not required in every implementation). Under the cumulative acknowledgement scheme, an acknowledgement is sent out for every other packet received to optimize bandwidth. Thus, in block 1212, a delayed acknowledgement timer is started if it has not been started already.

89 On the other hand, in block 1214, if the delayed acknowledgement timer has already started, the method stops the timer and proceeds to send out an acknowledgement to the other transmitting host computer. In any event, if the delayed acknowledgement timer expires, an acknowledgement is sent out to the transmitting host computer, in which case the acknowledgement only pertains to the one (instead of two) packets received. Generally speaking, the delayed acknowledgement timer may be set to a given value, e.g., 100ms. When the delayed acknowledgment timer expires, a request is then queued in the Rx NPP queue.

90 In block 1216, the method sends the data necessary to create an acknowledgement packet to the transmit network protocol processor (TX NPP) of the host computer to send out an acknowledgement. This data includes, for example, the acknowledgement number, the window size, and the pointer to the transmit control block (Tx TCB). This procedure has been described earlier in connection with Fig. 9 herein. The Rx NPP also has the option to send out an acknowledgement by itself. For example, the Rx NPP may send an immediate acknowledgement if it receives out-of-order data, duplicate data, or Rx window probe.

91 In block 1218, the method checks to determine whether the newly received data packets made the earlier partially reordered packets, which are stored in the reordered buffers, in order. If not, the process returns in block 1218. On the other hand, if the receipt of the newly received data packets made the earlier partially reordered packets in order, the method proceeds to block 1204 again to send out all the ordered packets to the host application buffer space and to send out acknowledgement for the same.

92 Fig. 12B represents the situation in Fig. 12A wherein the received packet is found to be out of order (as determined by block 1202 of Fig. 12A). In block 1250, the received packet is examined to determine whether it is an acknowledgement from the other transmitting host computer for data previously sent by this host computer or it is data, other than an acknowledgement, sent by the other transmitting host computer. If the received packet is found to be an acknowledgement, the method proceeds to step 1208 of Fig. 12A, i.e., the acknowledgement data is sent to the transmit network protocol processor (TX NPP) in order to allow the transmit network

protocol processor (TX NPP) to, among other tasks, update its window, calculate performance data, and the like.

93 On the other hand, if the received packet is found to be out of order data, other than an acknowledgement, sent from the other transmitting host computer, the method proceeds to block 1252 to determine whether the received data fits within the receive window. The receive window may be currently narrow because the host computer received a chunk of data earlier and has not completely moved the data received earlier into the application buffer memory, and thus can accept only a limited amount of new data. If the data does not fit within the received window, the method proceeds to block 1254 wherein the received data is simply discarded. Thereafter, the method returns to wait for the arrival of new data. This is shown in block 1256.

94 On the other hand, if the received data fits within the window, the method proceeds to block 1216 of Fig. 12A, i.e., it sends the received data, along with the receive control block (Rx TCB) to have the acknowledgement packet prepared and sent out by the transmit network protocol processor (TX NPP). The acknowledgement, in this case, will indicate to the other transmitting computer that the packets are received but they are out of order.

95 Each ROB is employed to store parameters associated with an out-of-order packet. The parameters associated with each ROB includes the following: Pointer to the data (PKT_PTR 1104), Storage Transport Layer Header (STL 1106), Length of current packet (PKT_LEN 1108), and Sequence number of current packet (SEQ_NUM 1110).

96 Further, the received data is stored in the out-of-order buffer as shown in block 1258. As discussed earlier, this out-of-order buffer is managed by parameters such as PKT_PTR 1104, STL 1106, PKT_LEN 1108, and SEQ_NUM 1110 in a linked list of reorder buffers, which is maintained in the receive control block (Rx TCB) and extends into the memory of the host computer as needed via the use of the buffer extension pointer ROB EXTENSION PTR 1112. Thus, in block 1260, the parameters associated with the reorder buffers in the receive control block (Rx TCB) are updated. Note that the Storage Transport Layer Header (STL) parameter

associated with the reorder buffers is modified only if the protocol (indicated by PROTOCOL 1124 in the receive control block (Rx TCB)) indicates that the current protocol is Fiber Channel over IP (FC/IP) over TCP/IP. Thereafter, the method returns to wait for the arrival of new data. This is shown in block 1262.

97 The separation of the transmit network protocol processor (TX NPP) from the receive network protocol processor (RCV NPP), in addition to the separation of the transmit control block (Tx TCB) from the receive control block (Rx TCB), advantageously allows the host computer to communicate at a higher transmission bandwidth. As mentioned above, each of the transmit network protocol processor (TX NPP) and receive network protocol processor (RCV NPP) executes their own threads for handling a variety of requests. Between the transmit network protocol processor (TX NPP) and the receive network protocol processor (RCV NPP), there is a substantial amount of required coordination.

98 In accordance with one aspect of the present invention, the receive network protocol processor (RCV NPP) passes data onto the transmit network protocol processor (TX NPP) on at least two occasions to update the transmit control block (Tx TCB) and other parameters associated with the transmit process. These two occasions include the receipt of an acknowledgement from the other transmitting host computer, and the receipt of data sequence information, which allows the transmit network protocol processor (TX NPP) to send an acknowledgement to the other transmitting host computer.

99 In the exemplary embodiment illustrated by Fig. 13, the receipt of one or more acknowledgement packets from the other transmitting host computer causes the receive network protocol processor (RCV NPP) 1302 to queue data associated with the received acknowledgement packet(s) into a queue 1304. Transmit network protocol processor (TX NPP) 1306 takes queue elements from queue 1304 in order to update its transmit control block (Tx TCB), the window size, calculate performance data, update the retransmit buffer, and the like. The TX NPP may then update the RX window based on the received acknowledgement. Furthermore, it may also update the send ACK # and the window size.

100 Among the information sent by receive network protocol processor (RCV NPP) 1302 to transmit network protocol processor (TX NPP) 1306 are the acknowledgement number, the time stamp value reflecting the time the packet(s) for which the acknowledgement is sent is received by the other transmitting host computer (the time stamp value is employed by the transmit network protocol processor (TX NPP) to calculate the delay based on the current time value), information associated with selective acknowledgement blocks such as the start sequence number (START_SEQ) and the associated length, the window size information from the other host transmitting computer, and the pointer to the transmit control block (Tx TCB).

101 In the exemplary embodiment illustrated by Fig. 14, the receive network protocol processor (RCV NPP) sends information to the transmit network protocol processor (TX NPP) to send out an acknowledgement. This may occur because the delayed acknowledgement timer has expired or because two packets back-to-back has been received and an acknowledgement should be sent. The receive network protocol processor (RCV NPP) 1402 queues requests to send acknowledgement into a queue 1404. Transmit network protocol processor (TX NPP) 1406 takes queue elements from queue 1404 in order to computer parameters necessary to send out an acknowledgement and also to update the transmit control block (Tx TCB) (e.g., the latest ACK and window size sent).

102 Among the information sent by receive network protocol processor (RCV NPP) 1402 to transmit network protocol processor (TX NPP) 1406 are the sequence number and the length, which allows the transmit network protocol processor (TX NPP) to calculate the acknowledgement number. Other information includes the window size to be sent to the other transmitting host computer, and the pointer to the transmit control block (Tx TCB). This information taken off the top of queue 1404 will be forwarded by transmit network protocol processor (TX NPP) 1406 to the header preparation processor (HDR NPP) create an acknowledgement packet for sending to the other host transmitting computer.

103 With respect to partially reordering packets, the invention involves techniques and structure for partially reordering packets in both a TCB that handles both transmitting

and receiving as well as in a control block arrangement in which there is a separate transmit TCB and a separate receive TCB (as discussed above herein). It should be noted that although the discussion below pertains to one implementation of partially reordering packet in a TCB that handles both transmitting and receiving, embodiments of the invention herein also applies to a control block arrangement in which there is a separate transmit TCB (Tx TCB) and a separate receive TCB (Rx TCB) for each networked device.

- 104 Fig. 15 illustrates, in accordance with one embodiment of the present invention, a receive transmit control block (Rx TCB) 1502 data structure for facilitating receiving data at the target device. Together with the transmit TCB (Tx TCB), Rx TCB 1502 facilitates the receive process at a target device by managing various aspects of the data transfer, such as the numbers of packets that are in transit at any given time, also called window management, and also by managing packet queuing and timing.
- 105 Incoming packets, themselves, are not directly kept in the re-order buffer (ROB) 1504 of the Rx TCB 1502. Since packets can be of varying size, it may be more efficient to store ROB pointers ("pkt_ptr") 1506 to the packets in the ROB 1504, and actually store some or all the packets themselves in an external memory location. Along with each pointer 1506, the packet length ("PKT_LEN[10:0]") 1508 and the sequence number ("SEQ_NUM") 1510 are stored. Of course the size of the packet length 1508 may be varied according to implementations.
- 106 The Rx TCB 1502 also contains a ROB extension pointer ("ROB extension ptr") 1512 for packet overflow. That is, arriving packets whose ROB pointers 1506 cannot fit in the ROB 1504 are themselves stored in a memory location that is referenced by the ROB extension pointer 1512. It is, in essence, a pointer to a location that contains other pointers, which can in turn reference packets.
- 107 The Rx TCB stores ROB pointers in its ROB 1504 in sequence number order. For example, the ROB pointer with the smallest sequence number is stored at the top 1516 of the ROB 1504, and the ROB pointer with the largest sequence number is stored at the bottom 1522 of the ROB 1504, or is referenced in a memory location by the ROB extension pointer 1512. ROB pointers in this memory location may also be stored by

sequence number. As new packets arrive, the ROB is reordered as necessary to insure that the ROB pointers are kept in order by sequence number. Also, packets may be discarded when their ROB pointers can no longer fit into the ROB 1504, or with the ROB extension pointer 1512, after readjustment.

108 Fig. 16 illustrates, in accordance with one aspect of the present invention, a simplified diagram showing an Rx TCB whose out-of-order packets are partially reordered. The reordering is only partial since reordering is undertaken even if all of the packets being reordered are not sequentially numbered. For example, reordering packets occurs in the buffer even though their numbers are not contiguous. The TCB in the example of Fig. 16 contains three slots in the ROB 1624 for each of three ordered entries, a first entry 1604, and second entry 1606, and a third entry 1608. Of course the number of slots can be varied as desired. Each entry may include a pointer to the received packet stored in the ROB, the length of the packet, and the sequence number of the packet. The TCB may also contain a counter 1601 that keeps track of the number of out-of-order packets that have arrived at the target device.

109 Once the packet with the expected sequence number 1602 arrives at the target device, the Rx TCB forwards it, along with all sequentially contiguous packets found in the ROB, to the application. The TCB determines the last sequentially contiguous entry in the memory buffer 1624 by subtracting the out-of-order counter 1601 from the total amount of packets received and stored.

110 Suppose three packets are transmitted from the source device to the target device. The first transmitted packet has a sequence number of 100, the second transmitted packet has a sequence number of 1000, and the third transmitted packet has a sequence number of 2000.

111 The packets arrive, however, out of order. The first packet received is the third transmitted packet with a sequence number of 2000. The second packet received is still the second transmitted packet with a sequence number of 1000. And the third packet received is the first transmitted packet with a sequence number of 100.

112 Initially, at time =N (322), prior to the arrival of any packets, the ROB 1624 is empty. The expected sequence number 1602 is 100 which is also the sequence

number of the first packet to be sent from the source device, since the TCP/IP data transfer has not yet started.

113 At time $=N + 1$ (310), the third transmitted packet arrives with sequence number 2000. Since this does not match the expected sequence number 1602 of "100", the third transmitted packet is placed in memory, and a ROB pointer 1616 is stored in the ROB 1624 at the first entry 1604, along with its length and sequence number. The out-of-order counter 1601 changed to "1". That is, there is one entry, and it is not in proper order.

114 At time $=N + 2$ (312), the second transmitted packet has arrived with sequence number 1000. Since this also does not match the expected sequence number 100, the second transmitted packet is placed in the memory, and a ROB pointer 1618 is stored in the ROB 1624, along with its length and the sequence number. Since the second transmitted packet has a sequence number smaller than the previously arrived third transmitted packet, the ROB pointer entries in the ROB 1624 are reordered. The ROB pointer to the second transmitted packet is placed in the first entry 1604, while the ROB pointer to the third transmitted packet is moved to the second entry 1606. The out-of-order counter 1601 is now changed to "2". That is, there are two entries, and they are both out of order, albeit partially reordered in ROB 1624.

115 At time $=N + 3$ (314), the first transmitted packet finally arrives, with sequence number that matches the expected sequence number 1602 of "100". At this point, the TCB places the first transmitted packet in memory, stores a ROB pointer 1620 in the ROB 1624 at the first entry 1606, along with its length and the sequence number. And again, the ROB 1624 is reordered. The ROB pointer to the second transmitted packet is moved to the second entry 1606, and the ROB pointer to the third transmitted packet is moved again to the third entry 1608. The out-of-order counter 1601 is now "0", since there are three entries, but all are corrected ordered.

116 The TCB then forwards all three properly ordered packets by sequence number to the application on the target device, clears the ROB, and subsequently determines the next expected sequence number 1602.

117 In the context of the present invention, providing a ROB for arrived packets reduces the need to flood the network with retransmitted packets. Also, reordering the ROB pointers by sequence number eliminates the need to continuously re-inspect the sequence number of each packet, thereby reducing network latency for the application.

118 Fig. 17 illustrates, in accordance with another embodiment of the present invention, a simplified diagram showing an exemplary Rx TCB that includes the partial reordering feature for out-of-order packets. In the embodiment of Fig. 17, packets that are sequentially contiguous with the packet having the next expected sequence number are forwarded to the application in the target device, and packets that are not sequentially contiguous with those forwarded are kept in the buffer to be reordered with packets subsequently arrived. In the example of Fig. 17, the Rx TCB contains three slots in the ROB 1724 for each of three ordered entries, a first entry 1704, and second entry 1706, and a third entry 1708. Each entry comprises a pointer to the received packet stored in the ROB, the length of the packet, and the sequence number of the packet. As in the Fig. 16, the TCB also contains a counter 1701 that keeps track of the number of out-of-order packets that have arrived at the target device.

119 In this example, there are four packets to be transmitted from the source device to the target device. The first transmitted packet has a sequence number of 100, the second transmitted packet has a sequence number of 1000, and the third transmitted packet has a sequence number of 2000, and the fourth transmitted packet has a sequence number of 3000.

120 The packets arrive, however, out of order. The first packet received is the fourth transmitted packet with a sequence number of 3000. The second packet received is still the second transmitted packet with a sequence number of 1000. And the third packet received is the first transmitted packet with a sequence number of 100. The third transmitted packet has not arrived.

121 Initially, at time =N (422), prior to any packets arriving, the ROB 1724 is empty. The expected sequence number 1702 is 100 which is also the sequence number of the

first packet to be sent from the source device, since the TCP/IP data transfer has not yet started.

122 At time $=N+1$ (410), the fourth transmitted packet arrives with sequence number 3000. Since this does not match the expected sequence number 100, the fourth transmitted packet is placed memory, and a ROB pointer 1716 is stored in the memory buffer 1724 at the first entry 1704, along with its length and sequence number. The out-of-order counter 1701 is changed to "1". That is, there is one entry, and it is not in proper order.

123 At time $=N+2$ (412), the second transmitted packet has arrived with sequence number 1000. Since this also does not match the expected sequence number 100, the second transmitted packet is placed memory, and a ROB pointer 1718 is stored in the memory buffer 1724, along with its length and the sequence number. Since the second transmitted packet has a sequence number smaller than the previously arrived fourth transmitted packet, the ROB pointer to the second transmitted packet is placed in the first entry 1704, and the ROB pointer to the fourth transmitted packet is moved to the second entry 1706. The out-of-order counter 1701 is changed to "2", since there are two entries, and both are out of order.

124 At time $=N+4$ (414), the first transmitted packet finally arrives, with a sequence number that matches the expected sequence number 1702. At this point, the TCB places the first transmitted packet in memory, stores a ROB pointer 1720 in the ROB 1724 at the first entry 1706, along with its length and the sequence number. The ROB pointer to the second transmitted packet is moved to the second entry 1706, and the ROB pointer to the fourth transmitted packet is moved to the third entry 1708. The out-of-order counter 1701 is now changed back to "1", since there are three total entries, the first two are in proper order, but the last one is out-of-order.

125 The TCB then forwards the first two properly ordered entries, the first & second transmitted packets, by sequence number to the application on the target device. It then moves ROB pointer to the fourth transmitted packet to the first entry 1704. And the TCB subsequently determines the next expected sequence number 1702.

126 In the context of the present invention, partially reordering the ROB pointers, and forwarding the group of packets by sequence number to the application without continuous re-inspection, reduces network latency for the application, as in the prior art shown in Fig. 15

127 Fig. 18 illustrates, in accordance with another aspect of the present invention, a simplified diagram showing the look-ahead algorithm with regard to the partial reordering of out-of-order packets in a TCB 1802. The TCB displays a ROB 1824 that can hold a total of five ROB pointer entries.

128 The source device has sent six packets, with sequence numbers of 100, 1000, 2000, 3000, 4000, and 5000 respectively. The target device has received just three of the packets, with sequence numbers 2000, 3000, and 4000 respectively. The corresponding ROB pointers are stored by sequence number in the ROB 1824.

129 The TCB look-ahead algorithm predicts, based on the lengths and sequence numbers of the already arrived packets, the next expected sequence number 1808, and/or the maximum transport unit (MTU) the number of packets that have not arrived and allocate space therefor in the buffer. The MTU is typically determined during, for example, the initialization process (e.g., during discovery). Suppose, for example, that the MTU is 1,500 bytes. In this case, the look-ahead algorithm would predict that at least two packets will be required to transport the first 1,900 bytes (i.e., the smallest received sequence number of 2000 minus the expected sequence number of 100).

130 Accordingly, the look-ahead algorithm allocate the next two entries in the ROB for the expected two packets. Any packet arriving whose sequence number is not between the smallest received sequence number of 2,000 and the expected sequence number of 100 is simply discarded. For example, if the packet with sequence number 5,000 arrives next, it will be discarded because the remaining two entries have already been allocated. On the other hand, if the packet with the expected sequence number 100 or a packet with a sequence number 1,000 arrives next, that packet will be put into the ROB and be taken into consideration during re-ordering.

- 131 In the context of the present invention, the TCB look-ahead algorithm reduces the unnecessary reordering of ROB pointers. If the TCB believes that that a packet with a higher sequence number needs to be discarded in the future to make room for other expected packets, it will discard that packet upon its arrival at the target device. This thereby reduces the overall network latency for the application, as groups of correctly ordered packets can be forwarded to the application efficiently by reducing the amount of reordering steps that need to occur.

Appendix A: Another Embodiment of the Invention

- 132 This invention relates generally to protocols for sending and receiving data across one or more computer networks and more specifically to implementing in hardware the Transmission Control Protocol/Internet Protocol used to transport data between two computers.
- 133 The Transmission Control Protocol/Internet Protocol ("TCP/IP") is a combination of protocols which allows computers (or "hosts") to communicate information across one or more networks. Within TCP/IP, each protocol is assigned a specific task, and the protocols are arranged in a layered design typically referred to as a "stack", i.e., a higher level protocol uses a lower level protocol to perform its tasks. Between TCP and IP, TCP is the higher level protocol. Moreover, TCP/IP processing is typically performed in software. Although TCP and IP are two of a family of many different protocols which perform their own unique tasks, it has become traditional to refer to the whole family of protocols as TCP/IP simply because TCP and IP are the best known of these protocols.
- 134 The Internet Protocol ("IP") is the protocol that provides for an unreliable transmission of packets of 64K bytes of data called datagrams from one host to another. This protocol is unreliable because it does not care if a sent packet is lost, duplicated, or corrupted. The IP relies on higher level protocols to ensure a reliable transmission. The primary tasks of the IP are addressing, routing, and fragmentation. Another duty of the IP is to remove old packets from the network.

135 The IP uses the address carried in the internet header to determine the destination of data packets. Selection of the path for transmission is called routing. Routes are based upon static or dynamic tables. Fragmentation is needed when a block of data is too large to fit the network below IP. If a large block of data arrives at the IP level for transmission over the network, the IP divides the block into smaller fragments before sending them onto the network. When data fragments are received from the network, the IP reassembles the data fragments so as to correspond to the original block.

136 The Transfer Control Protocol ("TCP") is designed to provide a reliable connection between two hosts across different networks using the Internet Protocol. The data sent with TCP/IP is considered to be a stream of bytes, even though the data has to be split into smaller packets due to the underlying protocols. TCP must recover data that is damaged, lost, duplicated, or delivered out of order by the internet communication system. TCP also provides a way for the receiver to govern the amount of data sent by the sender, i.e., it provides flow or congestion control. Another responsibility of the TCP protocol is connection handling, wherein TCP manages a number of connections at the same time.

137 The unit transferred between two host machines using the TCP protocol software is called a segment. Segments are exchanged to establish connections, to transfer data, to send acknowledgments ("ACK") that data was received, to define window sizes, and to close connections. Each segment is divided into two parts, a header and data. The TCP header carries the identification and control information and contains all the necessary information to ensure safe transfer of the data, to handle flow control, and to handle connections. Fig. A1 illustrates a conventional TCP segment with the data and the elements of the header, and is described in detail below.

138 As seen in Fig. A1, the Source Port is a 16-bit number that identifies where the TCP segment originated from within the sending host, and the Destination Port is a 16-bit number that identifies where the TCP segment is destined in the receiving host. The Sequence Number is a 32-bit number identifying the current position of the first data byte in the segment within the entire byte stream for the TCP connection. The

Acknowledgement Number is a 32-bit number identifying the next data byte the sender expects from the receiver.

139 HLEN is the header length, and it is a 4-bit field that specifies the total TCP header length. The field identified as Reserved is a 6-bit field currently unused and reserved for future use. Code bits are single bit flags that notify either the sending host or the receiving host that certain variables are being implemented during the TCP session. The Window is a 16-bit number used by TCP for flow control in the form of a data transmission window size, and it tells the sender how much data the receiver is willing to accept. The Checksum is preferably a 16-bit value that is computed by the TCP sending host, based on the contents of the TCP header and data fields. This 32-bit checksum will be compared by the receiver with the value the receiver generates using the same computation. If the values match, the receiver is assured that the segment arrived intact.

140 Where it is necessary for a TCP sender to notify the receiver of urgent data that should be processed by the receiving application as soon as possible, the Urgent Pointer is used. It is a 16-bit field that tells the receiver when the last byte of urgent data in the segment ends. In order to provide additional functionality, several optional parameters may be used between a TCP sender and receiver. Because options may vary in size, it may be necessary to "pad" the TCP header with zeroes, using the Padding field, so that the segment ends on a 32-bit word boundary as defined by the standard. Although the Data field is not used in some circumstances (e.g., ACK segments with no data in the reverse direction), this variable length field carries the application data from the TCP sender to receiver.

141 Local Area Networks ("LAN"), Metropolitan Area Networks ("MAN"), or Wide Area Networks ("WAN") typically use TCP/IP to transfer data from one computer to another, and Gigabit Ethernet is widely used as the physical medium in these environments. The Gigabit Ethernet, and similar technologies, enable fast transmission of the IP data once put on the wire. However, the prior art implements TCP/IP exclusively in software and is thus unable to utilize the potential increase in transfer rate through implementation of faster TCP/IP protocol processing. Therefore,

what is needed is a method for implementing TCP/IP protocol in both software and hardware.

142 The present invention provides for a method for transmitting a stream of data, comprising a plurality of data segments, from a sending computer to a receiving computer using a Transmission Control Protocol, the method comprising the steps of: (a) initializing a plurality of variables in a data structure, the data structure for use in transmitting the data segments in the data stream from the sending computer to the receiving computer, the plurality of variables in the data structure including: (1) an initial sequence number; and (2) the total number of data segments to be transmitted; and (b) transmitting the data segments using a plurality of microcode driven processing engines. The processing engines preferably comprise a re-transmission engine in the sending computer for determining whether a transmitted segment should be re-sent, an acknowledgement engine in the receiving computer for determining whether a transmitted segment was received by the receiving computer and for transmitting an acknowledgement signal to the sending computer, and an out-of-sequence detector in the receiving computer for detecting whether any the transmitted data segment was received out of sequence, and if so, for re-sequencing the data segments into their correct order. The step of transmitting the data segments comprises: (1) determining a rate at which the data segments are to be sent from the sending computer to the receiving computer; (2) identifying one of the data segments to be sent, assigning a sequence number to the selected data segment, and sending the selected data segment to the receiving computer; (3) initializing a retransmission timer as a function of the time when the selected data segment was transmitted and calculating a time duration value for the retransmission timer; (4) determining whether the time duration of the retransmission timer has expired before receipt of the acknowledgement signal for the selected data segment from the receiving computer and if so, resending the data segment and returning to step (3); and (5) determining whether additional the data segments are to be sent to the receiving computer, and if so repeating steps (1)-(5) until all data segments are successfully transmitted.

143 An object of the present invention is to define a new data structure (TCB) for use in implementing TCP/IP protocol in hardware. A key advantage of the present invention is that it increases the data transfer rate in TCP/IP protocol by using the

TCB data structure to implement TCP/IP protocol through microcode driven processing engines contained on one or more integrated circuit chips, e.g., Application Specific Integrated Circuit (ASIC) chips, instead of implementing TCP/IP protocol entirely through software. Therefore, efficiencies are achieved in the transmission and any necessary re-transmission of data, the acknowledgement that data was received, and in flow and congestion control.

144 The foregoing aspects and attendant advantages of the present invention will become more readily apparent by reference to the following detailed description, when taken in conjunction with the accompanying drawings, wherein:

145 Fig. A1 is a block diagram illustrating the elements of a prior art TCP data segment;

146 Fig. A2a illustrates the layout in memory of a TCB data structure according to a preferred embodiment of the present invention;

147 Fig. A2b is a flow chart illustrating Reorder Buffer (ROB) update and processing;

148 Figs. A3a-A3g list the preferred transmit variables, receive variables, and re-order variables used in a plurality of algorithms needed to implement the TCP protocol;

149 Fig. A4 is a flow diagram illustrating the steps for transmitting a stream of data, comprising a plurality of data segments, from a sending computer to a receiving computer according to a preferred embodiment of the present invention;

150 Fig. A5 is a block diagram illustrating an ASIC having microcode driven processing engines for TCP/IP processing according to the present invention;

151 Fig. A6 illustrates how Fibre Channel frame to Gigabit Ethernet frame processing is performed in the ASIC of Fig. A5; and

152 Fig. A7 illustrates how Gigabit Ethernet frame to Fibre Channel frame processing is performed in the ASIC of Fig. A5.

153 The present invention increases the data transfer rate in TCP/IP protocol by defining a new data structure (TCB) that implements TCP/IP through software microcode driven processing engines. TCB preferably occupies a contiguous block of memory of predetermined size. Fig. A2 illustrates an exemplary layout in memory of the TCB data structure according to a preferred embodiment of the present invention. Each TCB has associated with it pointers to three contiguous blocks of memory that store a "transmit, pending & acknowledgement" pending queue (TXP), a "received data segments processing" pending queue (RXP), and a buffer for reordering data segments if they are received out of sequence.

154 In addition, there are two queues called the transmit queue (TX) and the receive queue (RX) that are shared by all the TCP sessions. The transmit queue stores pointers to data that is to be transmitted to the receiving host and the RX queue stores pointers to data that has been received by the receiving host and is awaiting processing in memory. The Transmit Queue contains: a pointer that is maintained to point to the Transmit Pending Queue; a Write Index – TxQWrIndex; a Read Index – TxQRdIndex; and Retransmission Index – TxQRtxIndex. Finally, the Receive Queue contains a pointer that is maintained to points to the Start of the Receive Queue.

155 The Transmit Pending Queue is a contiguous block of memory of whose size determines the maximum number of pending entries that can be handled. This queue is maintained as a circular queue. Each entry consists of a pointer to the actual data segment, total length of the data segment, and Starting Sequence Number. Since memory is best utilized if the data is read in 64 byte chunks, each 64 byte chunk contains as many as four entries. To access this queue, three pointers identified as RdIndex, RtxIndex and WrtIndex, are maintained. Initially, these pointers are pointing to zero.

156 The functionality of the TXP is described by the following steps:

157 1. When a data block is received from the FC layer and is meant to be transmitted, it is put in the Transmit Pending Queue position given by the WrIndex. Update the WrIndex afterward.

158 2. When the window size of the TCP session is sufficient to transmit 1 or
more data segments, the data segments are picked up starting from the RdIndex
position and put into the TX queue. Update the RdIndex after that.

159 3. When an acknowledgement is received, then increment the RtxIndex
pointer beyond those data segments which have been covered by the
acknowledgement.

160 4. Whenever there is a timeout experienced because of a missing
acknowledgement, put the data segment indicated by the RtxIndex pointer into the TX
queue and reinitialize the timer with the next RTT value. (Exponential backoff
scheme may be used each time a retransmission is performed for the same data
segment.)

161 The functionality of the RXP is described by the following steps:

162 1. Any data segment on the RX queue that is identified to be meant for a
particular TCB should be put into the receive pending queue (RXP) of that particular
TCB. The receive sequence variables should be updated if the data received was in-
order.

163 2. The items in the receive pending queue store data for this particular
TCP session, but they may not be in order. Also some of the data segment may
contain only part of the FC data block since TCP may have broken up the FC data
block.

164 3. Reordering of items in the receive pending queue has to be done. To
reorder packets, the packets received are stored in a linked list form. Sequence
Number is used as the Key. When packets come out of order, they are stored in
ascending order of Sequence Numbers. Fig. A2b is a flow chart that illustrates
Reorder Buffer (ROB) update and processing according to the present invention.

165 4. Whenever more data is obtained in proper sequence, an ACK must be
generated to acknowledge the additional data received.

166 5. The new window has to be calculated depending on how much of the received data is sitting in memory and how much has already been transferred to the FC end. Thus, whenever an FC data block is transmitted to the receiver, this information should be used to update the window for TCP to release on the next transmission of an ACK. These contiguous blocks of memory could be allocated as and when the TCB gets used to set up an actual TCP session.

167 Figs. A3a-A3g list the preferred transmit variables, receive variables, and re-order variables used in a plurality of algorithms needed to implement the TCP protocol, and Fig. A4 is a flow diagram illustrating a process 40 for transmitting a stream of data, comprising a plurality of data segments, from a sending computer to a receiving computer according to a preferred embodiment of the present invention.

168 As illustrated in Figs A3a-A3g, there are three timers, the Retransmission Timer, the Idle Timer and the Persist Timer. Whenever the Retransmission Timer runs out, it should put an entry of a pointer to the TCB into the tasks queue for the retransmission process to work on. The Idle Timer is used to check for idle session and, if the timer runs out, then a dummy packet is sent to keep the session alive.

169 The Round Trip TCB Variables needed to calculate the Round Trip Delay include: a Current Retransmission Value-16 bit value (rxtcur); Log(2) of exponential backoff-16 bit value (rxtshift); a Round Trip time-16 bit value (rtt); a Minimum RTT value-16 bit value, which is calculated from the route characteristics (rttmin); a Current RT Sequence Number – 32 bit value (rtseq); a Smoothed Round Trip time-16 bit value – (srtt); and a Variance in Round Trip – 16 bit value (rttvar).

170 The TCP and IP Header parameters include: an IPDA (32 bit value) IP Destination Address; an IP Diffserv (8 bits) that specifies the Type of Service; and source and destination port values (16 bit values) to store the TCP source and destination port numbers.

171 Send Sequence Variables include: Send Unacknowledged (32 bit value) indicating the oldest unacknowledged sequence number; Send Next (32 bit value) indicating the sequence number that will be used for the next transmission; Send Urgent Pointer (32 bit value) indicating the 16-bit offset in the data where the urgent

data is sitting, which may be needed to communicate the other end that there is some urgent data in the packet transmitted; Send Window (16-bit value); Send Maximum-Highest Sequence Number Sent (32-bit value), which will be used for retransmission; and an Initial Sequence Number (32 bit value). The Receive Sequence Variables include: a Receive Window (16 bit value); a Receive Next (32 bit value); a Receive Urgent Pointer (32 bit value); and an Initial Receive Sequence Number (32 bit value).

172 When the receiving computer wants to send a stream of data to the receiving computer according to process 40 illustrated in Fig. A4., it establishes a TCP session. Then at step 41, the TCB Variables are initialized by software, preferably in the sending computer, in accordance with what is indicated in Figs. A3a-A3g. For instance, the Valid Flag variable indicates whether this is a valid TCB, and when a TCB is assigned this flag will be set to 1. However, care must be taken to clear certain fields, which might be there from a previous connection. Moreover, a Connection State variable reflects the current state of the TCP layer, and can be one of the following states: STATE_CLOSED; STATE_LISTEN; STATE_SYN_SENT; STATE_SYN_RECEIVED; STATE_ESTABLISHED; STATE_CLOSE_WAIT; STATE_FIN_WAIT_1; STATE_CLOSING; STATE_LAST_ACK; STATE_FIN_WAIT_2; STATE_TIME_WAIT. The Connection State is modified by the TCP layer and will be read by all the other blocks to figure out the state of a particular TCP session. Other variables initialized are those regarding an initial sequence number for the segments being transmitted and the total number of segments being transmitted.

173 Referring again to Fig. A4, once the TCB data structure is initialized, at step 42, the data segments in the data stream are transmitted using a plurality of microcode driven processing engines, according to steps 43-48. The microcode driven engines comprise a re-transmission engine in said sending computer for determining whether a transmitted segment should be re-sent, an acknowledgement engine in said receiving computer for determining whether a transmitted segment was received by said receiving computer and for transmitting an acknowledgement signal to said sending computer, and an out-of-sequence detector in said receiving computer for detecting whether any said transmitted data segment was received out of sequence, and if so, for re-sequencing said data segments into their correct order.

174 A rate at which the data segments are to be sent is determined, at step 43. The congestion control engine in used in determining this rate and is one of the microcode driven engines, and the receive sequence variable described above are calculated and used. The entire congestion control scheme is implemented in Microcode. User can configure this according to the environment. Current implementation is as follows: when RTO expires the congestion window is converted into 2*MSS (Maximum Segment Size). The recovery is done using a slow start process where in the congestion window is doubled every time an ACK is received. Duplicate ACK packets will be used to change the Congestion window size to 2*MSS. The first segment is re-transmitted. All these processes are done in hardware in the microcode driven engine.

175 At step 44, a data segment is identified to be sent, a sequence number is assigned to the data segment and the data segment is sent. In this step the send sequence variables discussed above are calculated and used, and the transmit pending queue is utilized. For instance, whenever a new data block arrives from a lower protocol layer, one or more nodes are added to the transmit pending queue to allow data transmission. Adding a Node consists of these steps:

- 176 1. If $(WrtIndex+1)\%MAX_TX_PENDING_ENTRIES == RtxIndex$, then this is over flow condition;
- 177 2. Memory Offset to Write = Queue Base Pointer + $((WrtIndex*16)/64)*64$;
- 178 3. Do a 64-byte block read using the above offset;
- 179 4. Add Offset = $((WrtIndex*16)\%64)$;
- 180 5. Now pointer to the data block and the total length can be added using the Add Offset;
- 181 6. Write back the 64-byte block using the "Memory Offset to Write";
- 182 7. Increment $WrtIndex = (WrtIndex + 1)\%MAX_TX_PENDING_ENTRIES$.

183 Deleting a node consists of these steps:

- 184 1. If $RdIndex == WrtIndex$, then the block is empty, nothing to read;
- 185 2. Memory Offset to Read = Queue Base Pointer +
(($RdIndex * 16$)/64)*64;
- 186 3. Do a 64-byte block read using the above offset;
- 187 4. Read Offset = ($RdIndex * 16$)%64;
- 188 5. Read the data block pointer and the total length;
- 189 6. If the action is delete this node, then $RdIndex$ has to be changed;
- 190 7. $RdIndex = (RdIndex + 1) \% MAX_TX_PENDING_ENTRIES$.

191 Once a segment has been sent, the retransmission engine is used to calculate the retransmission variables and set a retransmission timer at step 45. The retransmission algorithm for the retransmission timer is as follows. Every time a segment is sent, the TCP Engine starts a timer and waits for an acknowledgement. If the timer expires before acknowledgement is received, TCP Engine assumes the segment was lost or corrupted and retransmits it. The re-transmission timer value is calculated as

192
$$SRTT = 0.9 * OLD_SRTT + 0.1 * MRTT$$

193
$$RTO = 2 * SRTT$$

194 Where $RTO \rightarrow$ Retransmission Time out, $SRTT \rightarrow$ Smoothed Round Trip Time,

195 $MRTT \rightarrow$ Measured Round Trip Time.

196 All of these calculations are done in Protocol Processing Engines in Hardware. At step 46, it is determined whether the retransmission time has expired before an ACK is received on the selected segment, and if so, the segment is resent at step 47 and steps 45 and 46 are repeated. If not, then the ACK was sent. Moreover, it is

determined whether all data segments in the data stream were sent at step 43, and if not steps 43-48 are repeated until all data segments are successfully sent.

197 Date segment acknowledgement functions are performed in a microcode driven processing engine. Whenever an ACK is received for the TCB, the send window parameters change, and so additional window space may become available. When an ACK is received then Window is forwarded appropriately. Every time an ACK packet is received a sliding window (described in detail below) is moved and next set of packets is sent. Upon receiving segments properly (if the check sum is OK), an ACK packet is transmitted with an updated sequence number. If the receive buffer is full an ACK packet with window size 0 is sent.

198 TXP and ACK respectively keep track of where the next data segment to be transmitted is to be stored in memory, which is the next data segment to be put in the transmit queue (TX) and which is the next data segment to be retried in case the TCP session experiences a timeout while waiting for an ACK. The following actions are taken on an ACK reception.

199 1. When an ACK is received some data segments will have to be removed from the Transmit Pending Queue by moving the RtxIndex beyond all those data segments that have been properly acknowledged.

200 2. The send sequence variables have to be properly updated using the highest sequence number acknowledged and using the newly indicated window size.

201 3. If there is data along with the ACK, the receive sequence variables should also be updated (if the data is in order), and the data segment should be stored in memory and a pointer to the same should be added into the RXP queue of the TCB.

202 Through the transmission of data segments, one or more segments may be received out of sequence. The out of sequence detector is the microcode driven processing engine used in such a situation. For instance, let us assume a seven segments with their byte streams as follows

203

	Seg 2, 1400	Seg 3, 2000	Seg 4, 2400	Seg 5, 2800	Seg 6, 3200	Seg 7, 4000
--	----------------	----------------	----------------	----------------	----------------	----------------

204 Segment 1 arrives at the receiver which acknowledges it by sending an ACK segment with ACK=1000 and WIN=1600 (Window size 1600). Since ACK+WIN=2600 the sender can send segments 2,3 and 4. Segments 2 and 3 are sent. Segment 3 has arrived but segment 2 has been delayed. The receiver sends another ACK segment with ACK=1000 and WIN=1600. The sender sends segment 4. Segment 4 arrives but segment 2 is still outstanding. Again the receiver sends ACK=1000, WIN=1600. However the sender cannot send segment 5 since it would take the sent sequence number to 2800 which is greater than ACK+WIN. The sender cannot do anything further unless Segment 5 is split into smaller segments to use up the remaining part of the window. Segment 2 reaches the receiver and the receiver announces a new ACK value. The delayed segment 2 now reaches the receiver which sends ACK=2400 and WIN=1600 allowing the sender to send up to 4000.

205 The initial values of the sequence numbers are exchanged during the connection establishment sequence. This flow control mechanism shown above is usually called a sliding window protocol. When packets are received out of Sequence it is stored in the Re-order buffer. These packets are sorted and stored in the external memory.

206 The Hardware implementation of TCP Engine is done using a set of protocol processing engines, which are microcode driven. Initial parameters for processing are setup by the software through an external CPU. Calculations for the algorithms are done using the above described microcode driven engines. This method provides flexibility for changes.

207 Fig. A5 is a block diagram illustrating an ASIC having microcode driven processing engines for TCP/IP processing according to the present invention;

208 Fig. A6 illustrates how Fibre Channel frame to Gigabit Ethernet frame processing is performed in the ASIC of Fig. A5; and

209 Fig. A7 illustrates how Gigabit Ethernet frame to Fibre Channel frame processing is performed in the ASIC of Fig. A5.

210 Thus, the invention relates, in one embodiment to a method for transmitting a stream of data, comprising a plurality of data segments, from a sending computer to a receiving computer using a Transmission Control Protocol, the method comprising the steps of initializing a plurality of variables in a data structure, the data structure for use in transmitting the data segments in the data stream from the sending computer to the receiving computer, the plurality of variables in the data structure including an initial sequence number; and the total number of data segments to be transmitted. The method further includes transmitting the data segments using a plurality of microcode driven processing engines, the processing engines comprising, a re-transmission engine in the sending computer for determining whether a transmitted segment should be re-sent, an acknowledgement engine in the receiving computer for determining whether a transmitted segment was received by the receiving computer and for transmitting an acknowledgement signal to the sending computer, a congestion control engine for determining a rate at which data segments are to be sent from the sending computer to the receiving computer, and an out-of-sequence detector in the receiving computer for detecting whether any the transmitted data segment was received out of sequence, and if so, for re-sequencing the data segments into their correct order. The step of transmitting the data segments comprises (1) determining the rate at which the data segments are to be sent from the sending computer to the receiving computer, (2) identifying one of the data segments to be sent, assigning a sequence number to the selected data segment, and sending the selected data segment to the receiving computer, (3) initializing a retransmission timer as a function of the time when the selected data segment was transmitted and calculating a time duration value for the retransmission timer, (4) determining whether the time duration of the retransmission timer has expired before receipt of the acknowledgement signal for the selected data segment from the receiving computer and if so, resending the data segment and returning to step (3) and (5) determining whether additional the data segments are to be sent to the receiving computer, and if so repeating steps (1)-(5) until all data segments are successfully transmitted.

211 In one embodiment, the microcode driven re-transmission engine comprises a re-transmission timer algorithm for calculating a re-transmission timer value, the

algorithm depending upon a smoothed round trip time value and a measured round trip time value.

212 In another embodiment, the microcode driven acknowledgement engine comprises at least one algorithm, the at least one algorithm configured to send the acknowledgement signal when the data segment is properly received by the receiving computer; and move a sliding window and instruct sending the next data segment in the stream of data when the sending computer receives the acknowledgement signal; and send the acknowledgement signal, wherein a window size header element is zero when a receive buffer is full.

213 In another embodiment, the microcode driven congestion control engine comprises a plurality of algorithms, the plurality of algorithms including a slow start algorithm for doubling the size of a congestion window when each the acknowledgement signal is received by the sending computer, an algorithm for changing the size of the congestion window when a re-transmission timer expires, the algorithm depending upon a maximum segment size, and an algorithm for changing the size of the congestion window when a duplicate acknowledgement signal is received by the sending computer, the algorithm depending upon the maximum segment size.

214 In another embodiment, the microcode driven out-of-sequence detector comprises a flow control algorithm for implementing a sliding window protocol.

215 In one embodiment, there is disclosed a method for transmitting a stream of data, having a plurality of data segments, from a sending computer to a receiving computer using a Transmission Control Protocol. The method includes the steps of initializing a plurality of variables in a data structure that will be used in transmitting the data segments and transmitting the data segments using a plurality of microcode driver processing engine that each perform an algorithm to process the initialized variables and to update those variables.

216 The implementation example of a method for implementing TCP/IP protocol in hardware described in the text above was chosen as being illustrative of the best mode of the present invention. The preferred embodiment of the present invention

described above is illustrative of the principles of the invention and is not intended to limit the invention to the particular embodiment described. Accordingly, while the preferred embodiment of the present invention has been illustrated and described, it will be appreciated that various changes can be made therein without departing from the spirit and scope of the invention.

217 Thus, while this invention has been described in terms of several preferred embodiments, there are alterations, permutations, and equivalents which fall within the scope of this invention. It should also be noted that there are many alternative ways of implementing the methods and apparatuses of the present invention. It is therefore intended that the following appended claims be interpreted as including all such alterations, permutations, and equivalents as fall within the true spirit and scope of the present invention.

CLAIMS

What is claimed is:

1. In a first host computer coupled to a network, a method for exchanging data with a second host computer also coupled to said network, said method comprising:

employing a first Tx Transmit Control Block (Tx TCB) to facilitate transmitting first data associated with a bi-directional data flow from said first host computer to said second host computer, said first data being transmitted using the TCP protocol, said first Tx TCB being associated with said first host computer and is configured for storing transmit-facilitating parameters associated with said transmitting said first data using said TCP protocol; and

employing a first Rx Transmit Control Block (Rx TCB) to facilitate receiving second data associated with said bi-directional data flow, said second data being transmitted from said second host computer to said first host computer using said TCP protocol, said first Rx TCB being associated with said first host computer and is configured for storing receive-facilitating parameters associated with said receiving said second data using said TCP protocol, wherein said first Tx TCB and said first Rx TCB are employed to simultaneously service a transmit request pertaining said first data and a receive request pertaining said second data.

2. Circuitries for facilitating data exchange via a network, said circuitries being associated with a first host computer coupled to said network, said network also being coupled to a second host computer, said circuitries comprising:

means for facilitating transmitting first data associated with a bi-directional data flow from said first host computer to said second host computer, said first data being transmitted using the TCP protocol, said means for facilitating transmitting being associated with said first host computer and is configured for storing transmit-facilitating parameters associated with transmitting said first data using said TCP protocol; and

means for facilitating receiving second data associated with said bi-directional data flow, said second data being transmitted from said second host computer to said

first host computer using said TCP protocol, said means for facilitating receiving being associated with said first host computer and is configured for storing receive-facilitating parameters associated with receiving said second data using said TCP protocol, wherein said means for facilitating transmitting and said means for facilitating receiving are employed simultaneously to facilitate simultaneous transmitting said first data and receiving said second data by said circuitries.

3. On a network interface card (NIC card) coupled to a first host computer, a method for exchanging data via a network coupled to said first host computer and a second host computer, said method comprising:

transmitting, using transmit-facilitating parameters stored in a first Tx Transmit Control Block (Tx TcB) associated with said NIC card, first data associated with a bi-directional data flow, said first data being transmitted from said first host computer to said second host computer, said first data being transmitted using the TCP protocol; and

receiving, using receive-facilitating parameters stored in a first Rx Transmit Control Block (Rx TcB) associated with said NIC card, second data associated with said bi-directional data flow, said second data being transmitted from said second host computer to said first host computer, said second data being transmitted using the TCP protocol, wherein said Tx TcB and said Rx TcB are employed to facilitate simultaneous transmitting said first data and receiving said second data by said NIC card.

4. A method for transmitting transmit data associated with a bi-directional data flow between a first host computer and a second host computer, said first host computer and said second host computer being coupled via a computer network, said method comprising:

ascertaining, from a first control block implemented in said first host computer and associated with said bi-directional data flow, a transmit window size parameter, said first control block being configured to track transmit-facilitating parameters employed for transmitting, using the TCP protocol, said transmit data associated with said bi-directional data flow; and

if said transmit window size parameter indicates that a transmit window is available to transmit at least a minimum segment size of said transmit data, tracking, using said first control block, a first portion of said transmit data in a retransmit queue, and preparing at least one TCP header for said first portion of said transmit data to facilitate transmitting of said first portion of said transmit data to said second host computer using said TCP protocol, said first portion of said transmit data representing an amount of transmit data that can be transmitted in view of said transmit window size parameter,

wherein said first host computer is also configured to receive receive data associated with said bi-directional data flow from said second host computer simultaneously with said transmitting said first portion of said transmit data, receive-facilitating parameters employed to facilitate receive of said receive data being tracked in a second control block at the same time that transmit-facilitating parameters employed to transmit said transmit data are tracked in said first control block, said second control block being implemented in said first host computer and associated with said bi-directional data flow.

5. A method for transmitting transmit data associated with a bi-directional data flow between a first host computer and a second host computer, said first host computer and said second host computer being coupled via a computer network, said method comprising:

storing transmit-facilitating parameters employed for said transmitting said transmit data in a first control block, said first control block being implemented in said first host computer and associated with said bi-directional data flow, said transmitting said transmit data being performed in accordance with the TCP protocol; and

employing said transmit-facilitating parameters in said first control block to facilitate servicing a transmit request pertaining to a given portion of said transmit data from said first host computer to said second computer, wherein said servicing said transmit request occurs simultaneously with servicing a receive request pertaining to receive data associated with said bi-directional data flow from said second host computer by said first host computer, said receive data being received using receive-

facilitating parameters stored in a second control block implemented in said first host computer, said receive data being received in accordance with said TCP protocol.

6. Circuitries for facilitating data exchange via a network, said circuitries being associated with a first host computer coupled to a network, said network being coupled to a second host computer, said circuitries comprising:

means for storing transmit-facilitating parameters employed for transmitting transmit data associated with a bi-directional data flow from said first host computer to said second host computer, said transmit data being transmitted using the TCP protocol, said means for storing being implemented in said first host computer; and

means for servicing a transmit request pertaining a given portion of said transmit data using said transmit-facilitating parameters, wherein said servicing said transmit request occurs simultaneously with servicing a receive request pertaining to receive data associated with said bi-directional data flow from said second host computer by said first host computer, said receive data being received using parameters tracked in means for storing receive-facilitating parameters, said means for storing receive-facilitating parameters also being implemented in said first host computer, said receive data being received in accordance with said TCP protocol.

7. Circuitries for facilitating data exchange via a network, said circuitries being associated with a first host computer coupled to said network, said network being coupled to a second host computer, said circuitries comprising:

a first control block configured to store transmit-facilitating parameters employed for transmitting transmit data associated with a bi-directional data flow from said first host computer to said second host computer, said transmit data being transmitted using the TCP protocol, said first control block being implemented in said first host computer; and

circuitry configured to service a transmit request pertaining a given portion of said transmit data using said transmit-facilitating parameters, wherein said servicing said transmit request occurs simultaneously with servicing a receive request pertaining to receive data associated with said bi-directional data flow from said second host computer by said first host computer, said receive data being received using parameters tracked in a second control block, said second control block also being

implemented in said first host computer, said receive data being received in accordance with said TCP protocol.

8. A method for receiving at a first host computer receive data associated with a bi-directional data flow between said first host computer and a second host computer, said first host computer and said second host computer being coupled via a computer network, said method comprising:

ascertaining, responsive to a receipt of a first set of packets received at said first host computer, an expected sequence number parameter, said first set of packets being associated with said bi-directional data flow, said expected sequence number parameter being stored in a first control block implemented in said first host computer and associated with said bi-directional data flow, said first control block being configured to store receive-facilitating parameters employed for receiving, using the TCP protocol, said receive data; and

if said expected sequence number parameter indicates that a first set of packets received at said first host computer represents packets in order, updating in said first control block said expected sequence parameter,

wherein said first host computer is also configured to transmit, using the TCP protocol, transmit data associated with said bi-directional data flow from said first host computer to said second computer, a request to transmit a portion of said transmit data being serviced simultaneously with receiving said first set of packets, transmit-facilitating parameters employed to transmit said transmit data being stored in a second control block at the same time that said receive-facilitating parameters are stored in said first control block, said second control block being implemented in said first host computer and associated with said bi-directional data flow.

9. A method for receiving receive data associated with a bi-directional data flow between a first host computer and a second host computer, said first host computer and said second host computer being coupled via a computer network, said method comprising:

storing receive-facilitating parameters employed for said receiving said receive data in a first control block, said first control block being implemented in said first

host computer and associated with said bi-directional data flow, said receiving said receive data being performed in accordance with the TCP protocol; and

employing said receive-facilitating parameters in said first control block to facilitate receiving a given portion of said receive data at said first host computer from said second computer, wherein said receiving said given portion of said receive data occurs simultaneously with servicing a transmit request pertaining to transmit data associated with said bi-directional data flow from said first host computer to said second host computer, said transmit data being transmitted using transmit-facilitating parameters stored in a second control block implemented in said first host computer, said transmit data being transmitted in accordance with said TCP protocol.

10. Circuitries for facilitating data exchange via a network, said circuitries being associated with a first host computer coupled to said network, said network being coupled to a second host computer, said circuitries comprising:

a first control block configured to store receive-facilitating parameters employed for receiving receive data associated with a bi-directional data flow at said first host computer from said second host computer, said receive data being received using the TCP protocol, said first control block being implemented in said first host computer; and

circuitry configured to receive a given portion of said receive data using said receive-facilitating parameters, wherein said receiving said given portion of said receive data occurs simultaneously with servicing a transmit request pertaining to transmit data associated with said bi-directional data flow from said first host computer to said second host computer, said transmit data being transmitted using parameters stored in a second control block, said second control block also being implemented in said first host computer, said transmit data being transmitted in accordance with said TCP protocol.

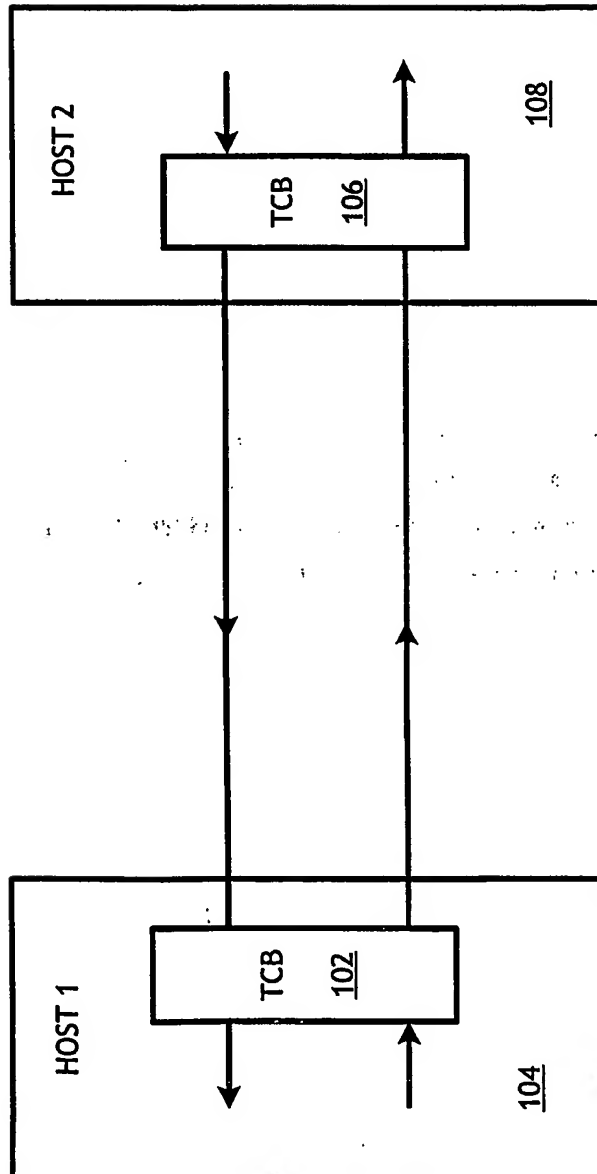


FIG. 1

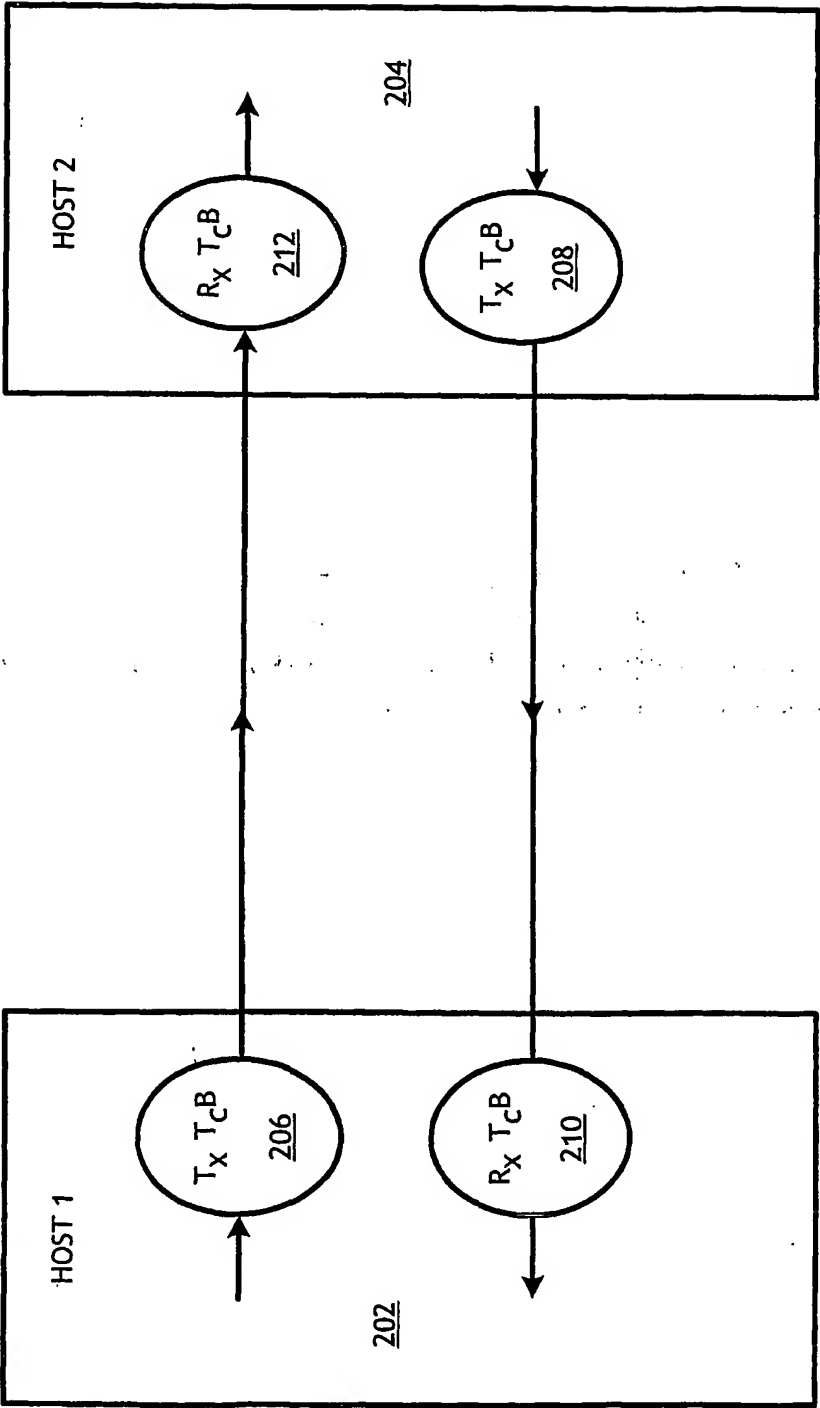


FIG. 2

302							312		314		316		318		T _X T _C B	
7		6		5		4		3		2		1		0		
TCP SP		TCP DP		304		IP DA								306		
308 TTL		310 TOS		PROTOCOL		UNUSED		SND_WIN[31:16]		SEG_OPT_MSS[15:0]						
OLD_UNA				320		SND_NXT								322		
ICNG_WIN[31:16], CNG_WIN[15:0]				324		SND_MAX								326		
MSS[31:16], MAX_WIN_SZ[15:0]				328		TXP_BASE_ADDR								330		
TMR_SEQ_NUM				332		RTO								334		
MRTT				336		SRTT								338		
FST				340		MAX_TXP_ADDR								342		
TXP_WR_PTR				344		TXP_RD_PTR								346		
TXP_RTX_PTR				348		TXP_SHD_RD_PTR								350		
DUP_ACK_CNT[60:56], M_HL[44:40], TCP_STATE[39:32], PER_CNTR[28:24], RTX_CNTR[20:16], MAX_PER_CNTR[4:0]														352		
PTO				356		TMR_SEQ_NUM_VAL[15], RTX_ENT_VAL[8], PER_ENT_VAL[0]										
						MAC HEADER										
						MAC HEADER										
						MAC HEADER										
IP SA				366		TCP_RTX_SEGS		370		RCV_WIN_SIZE		354		TCP_OUT_SEGS		372

FIG. 3

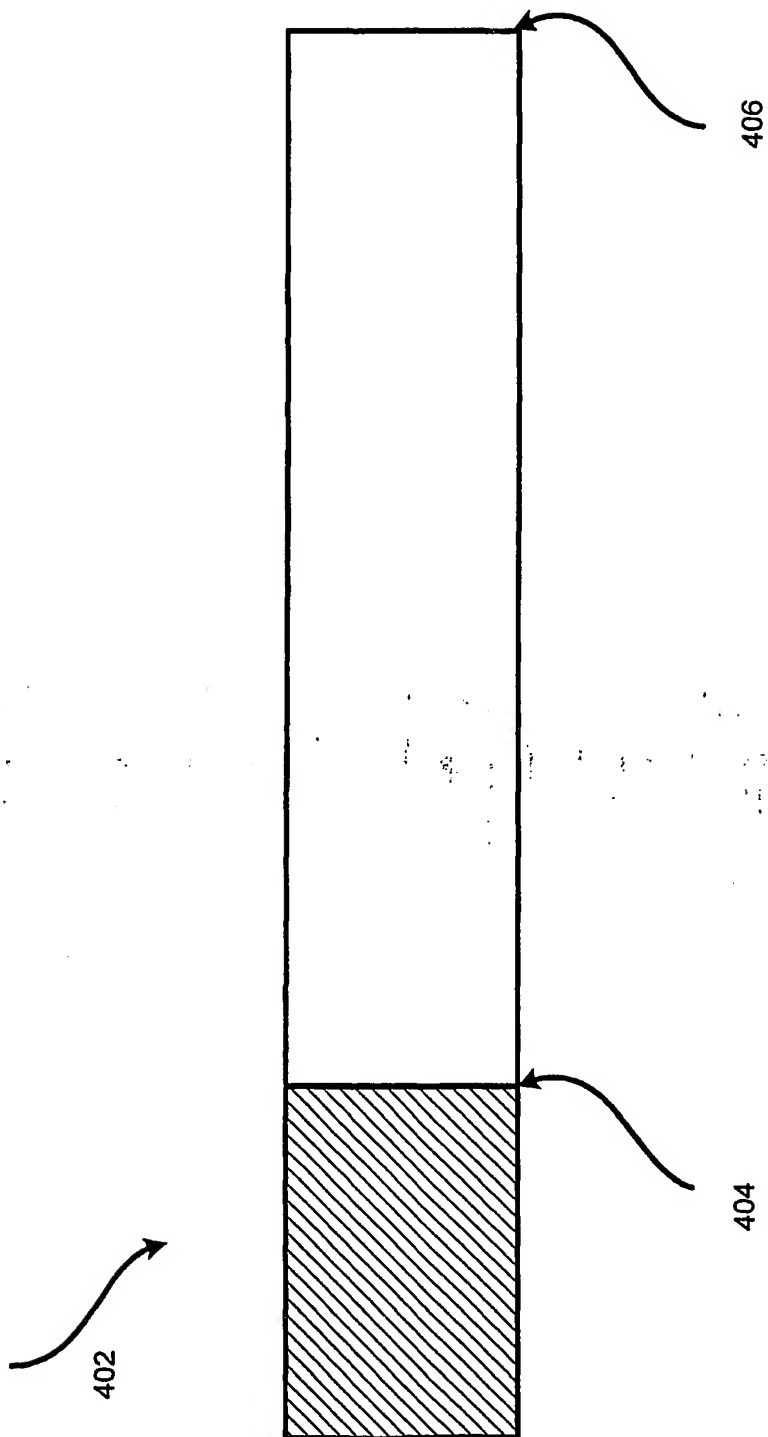


FIG. 4

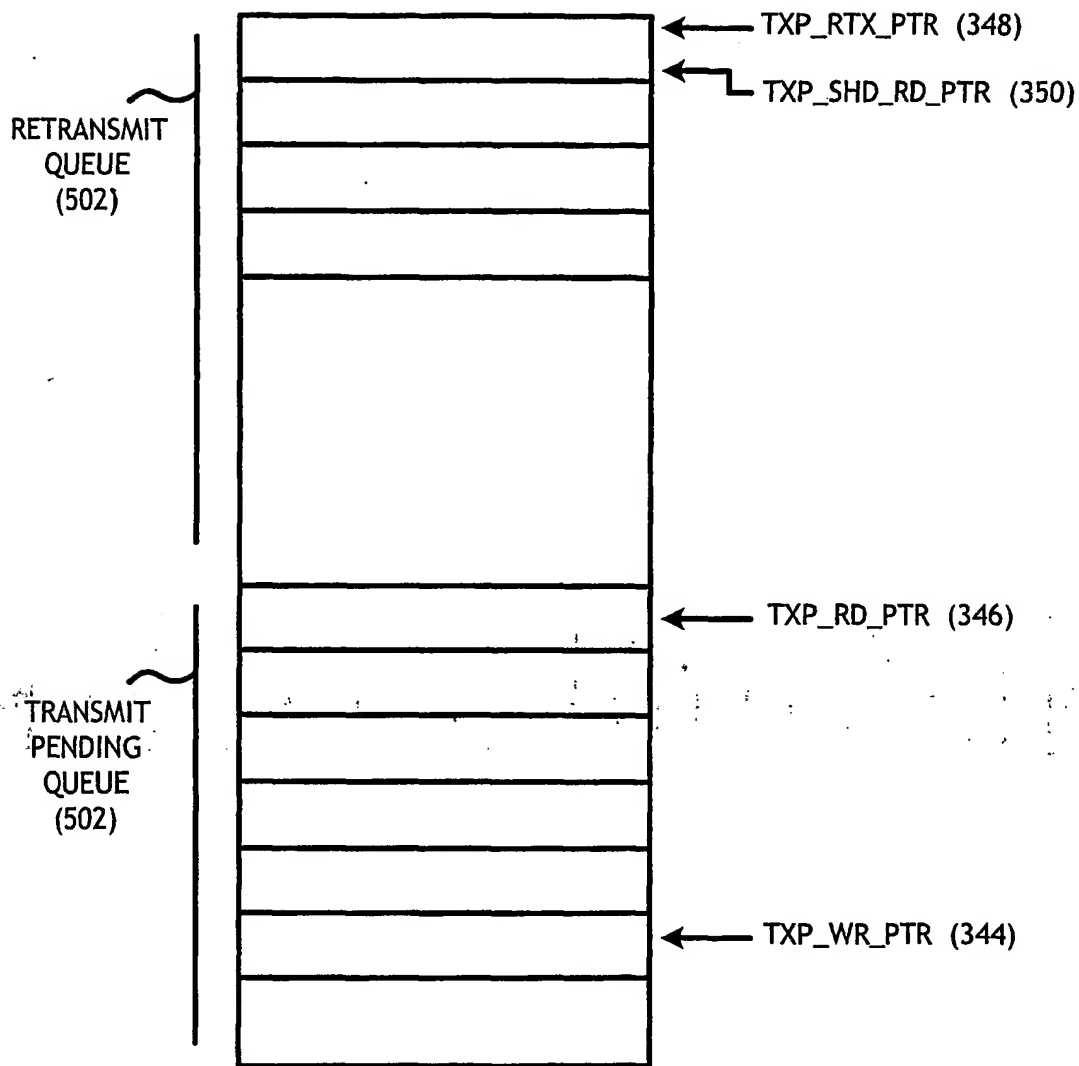
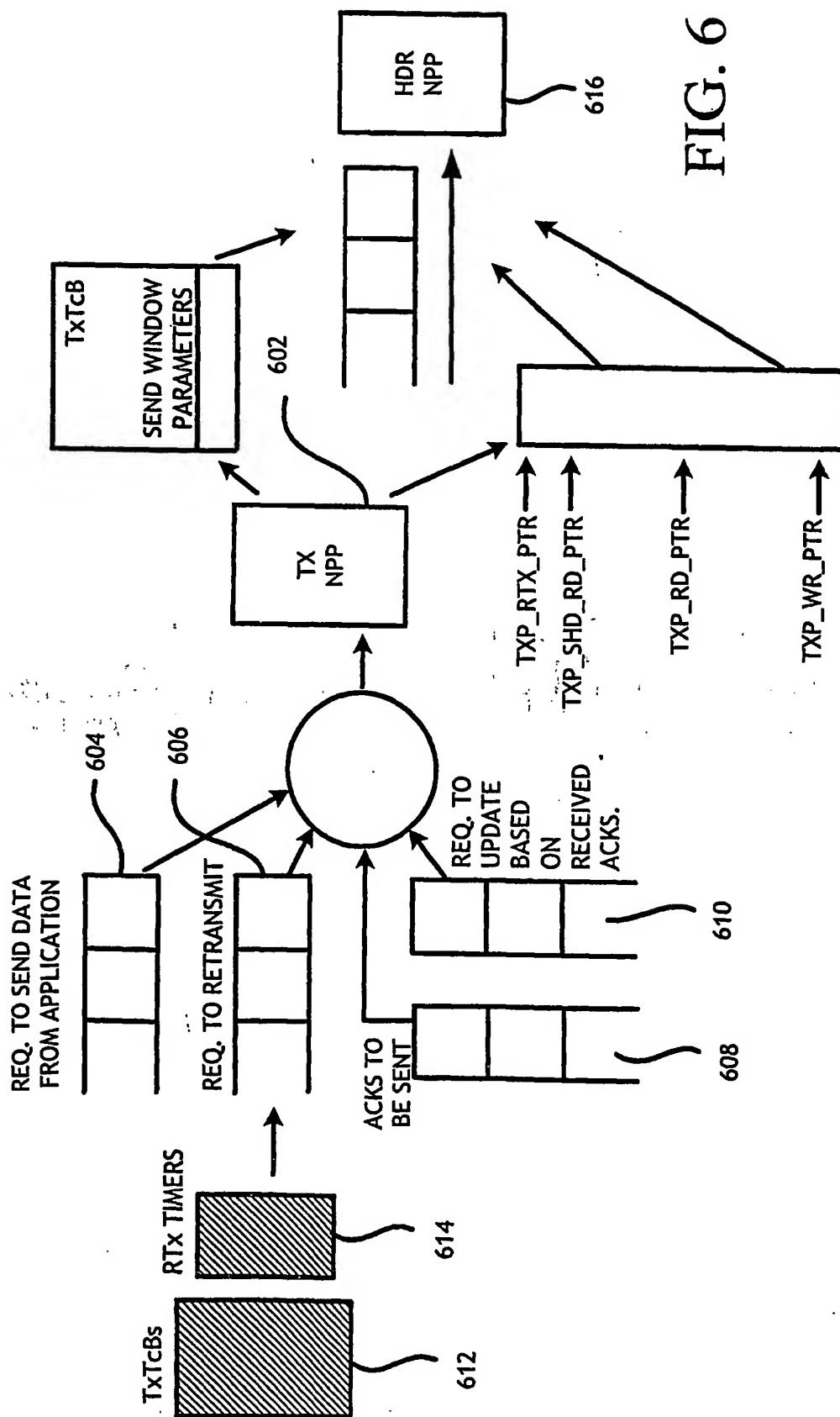


FIG. 5



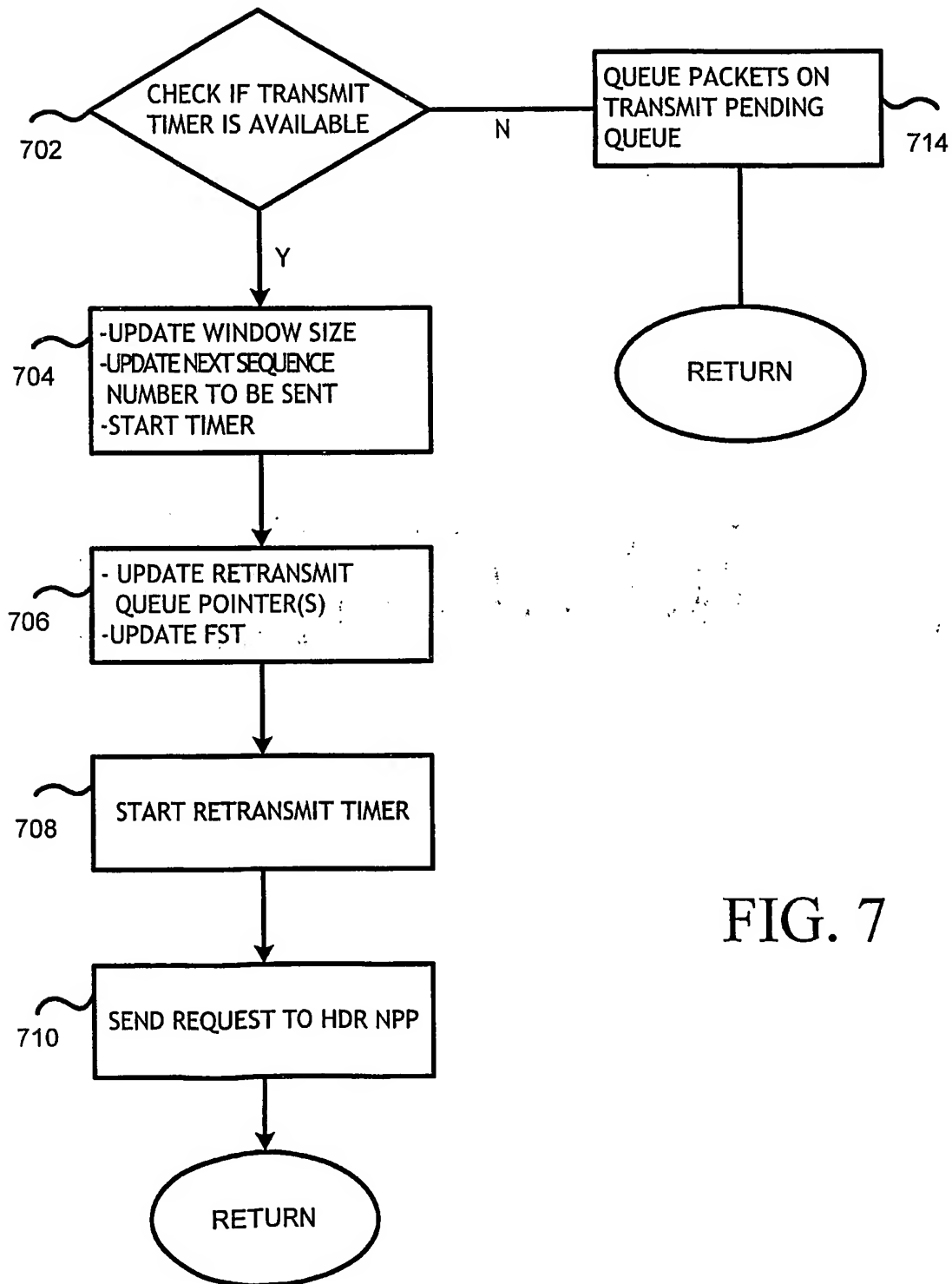


FIG. 7

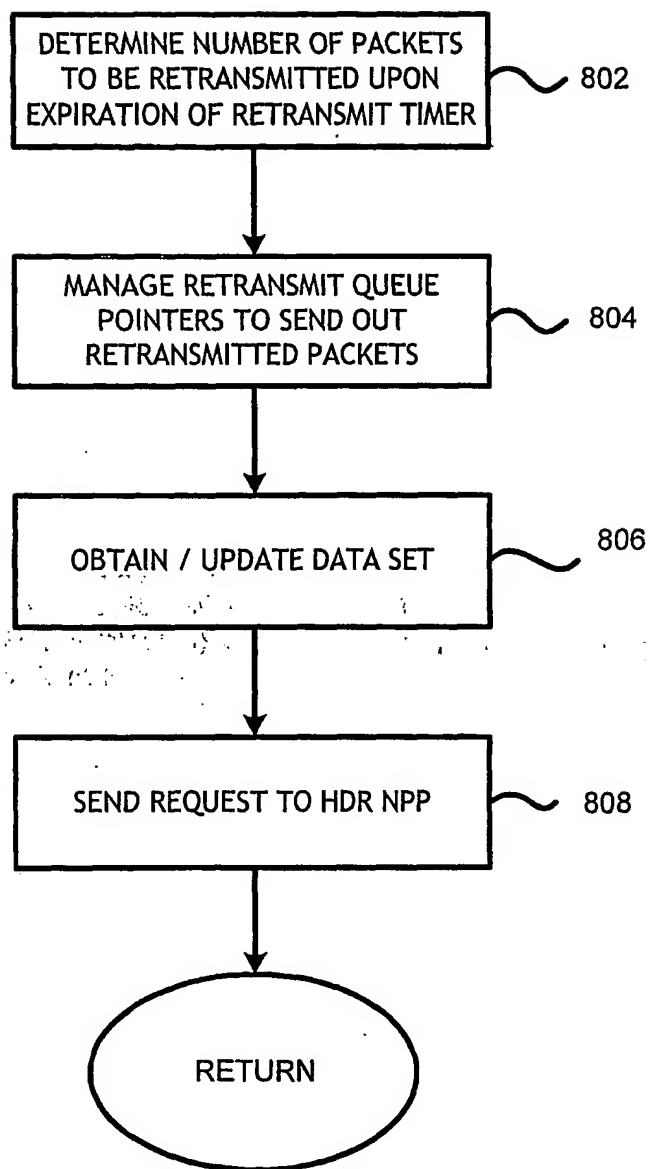


FIG. 8

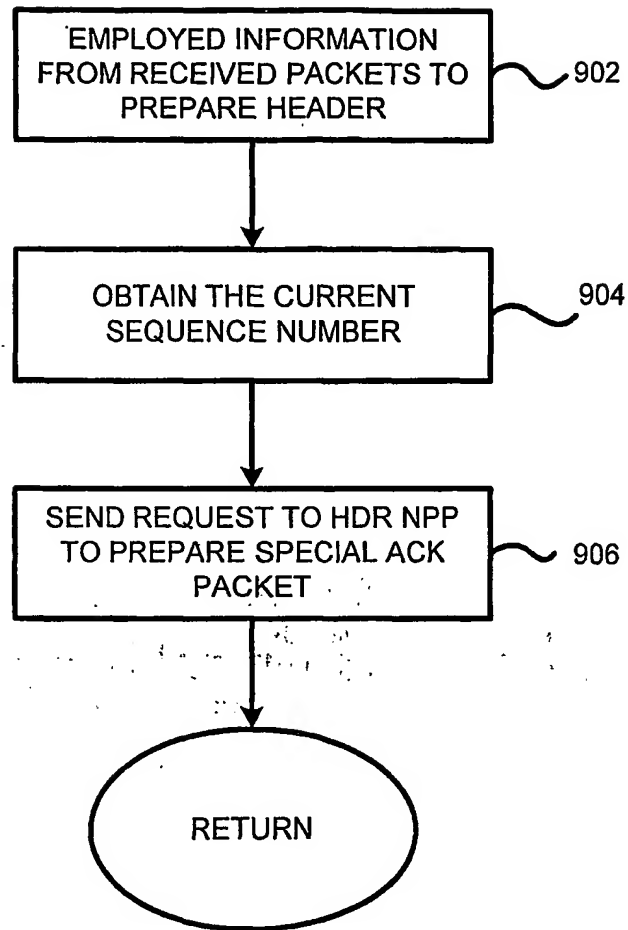


FIG. 9

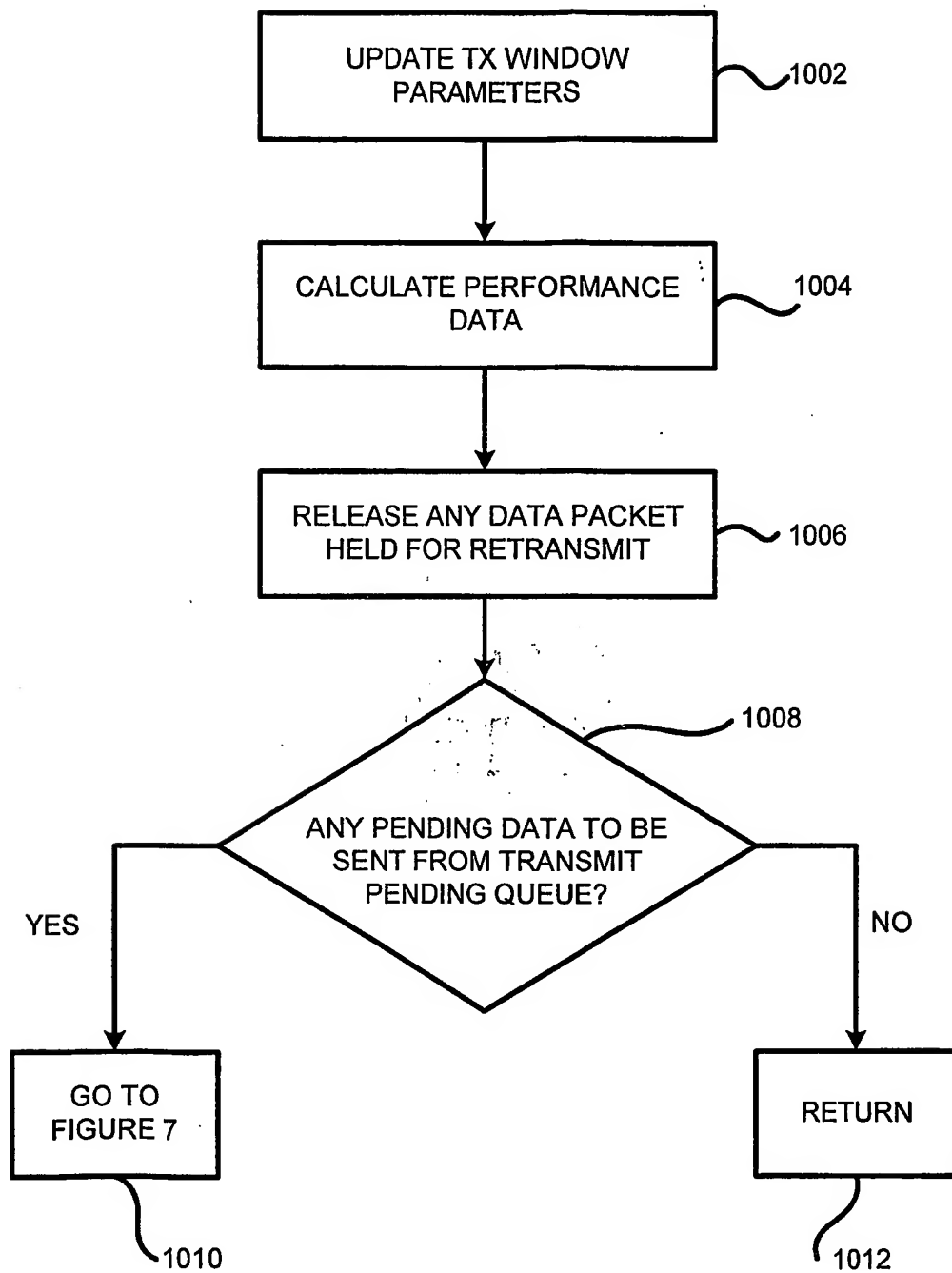


FIG. 10

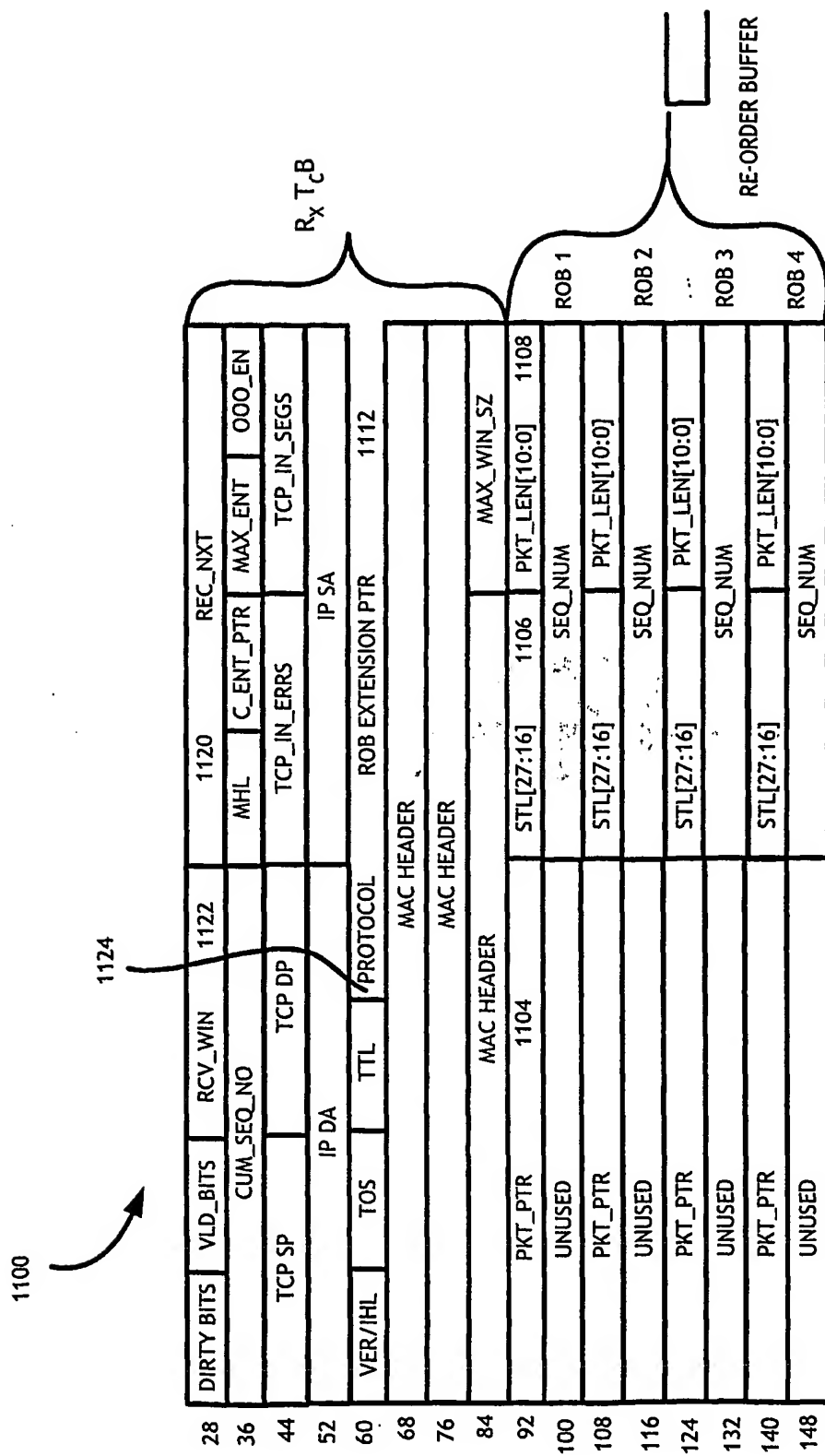
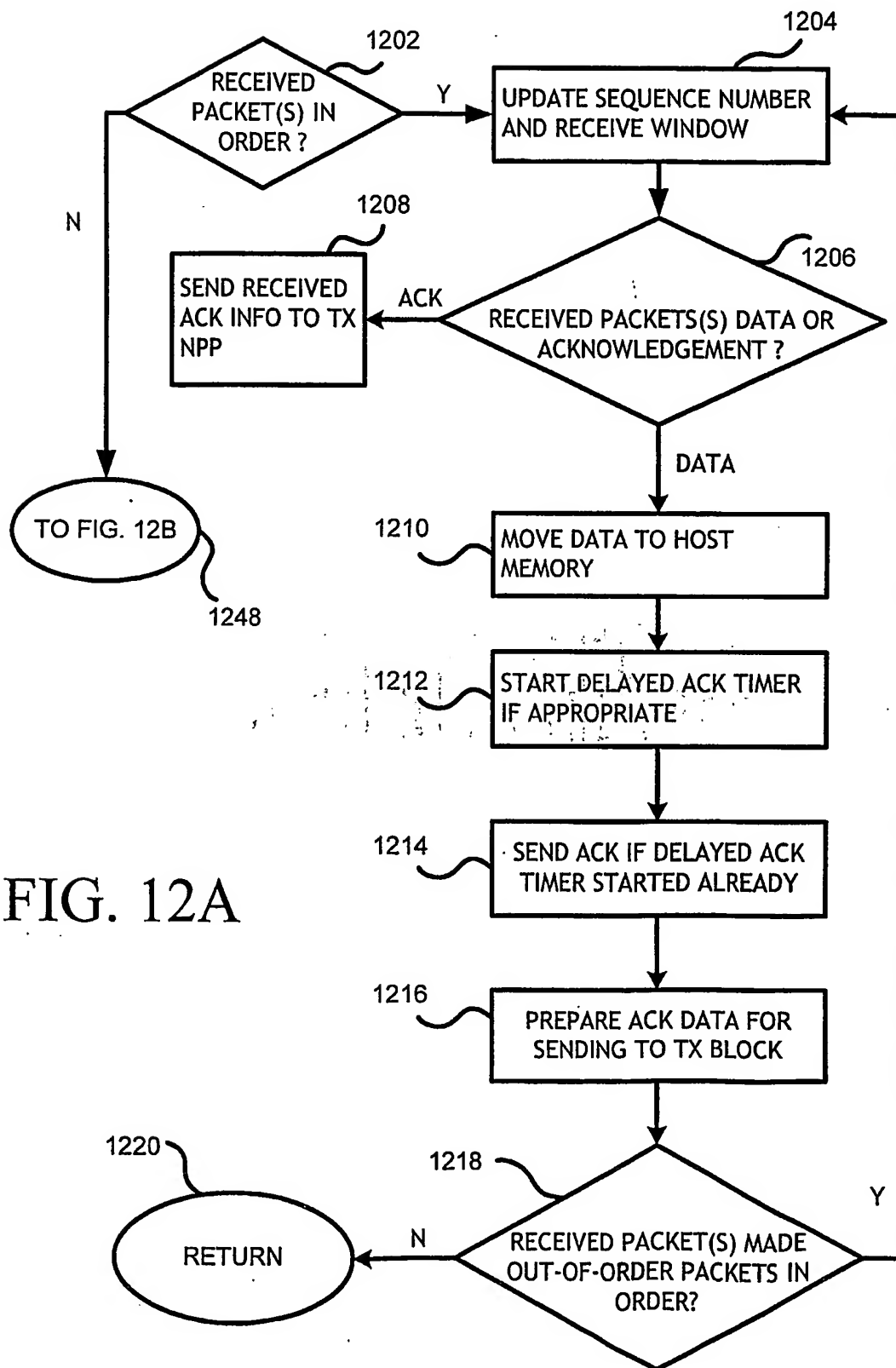
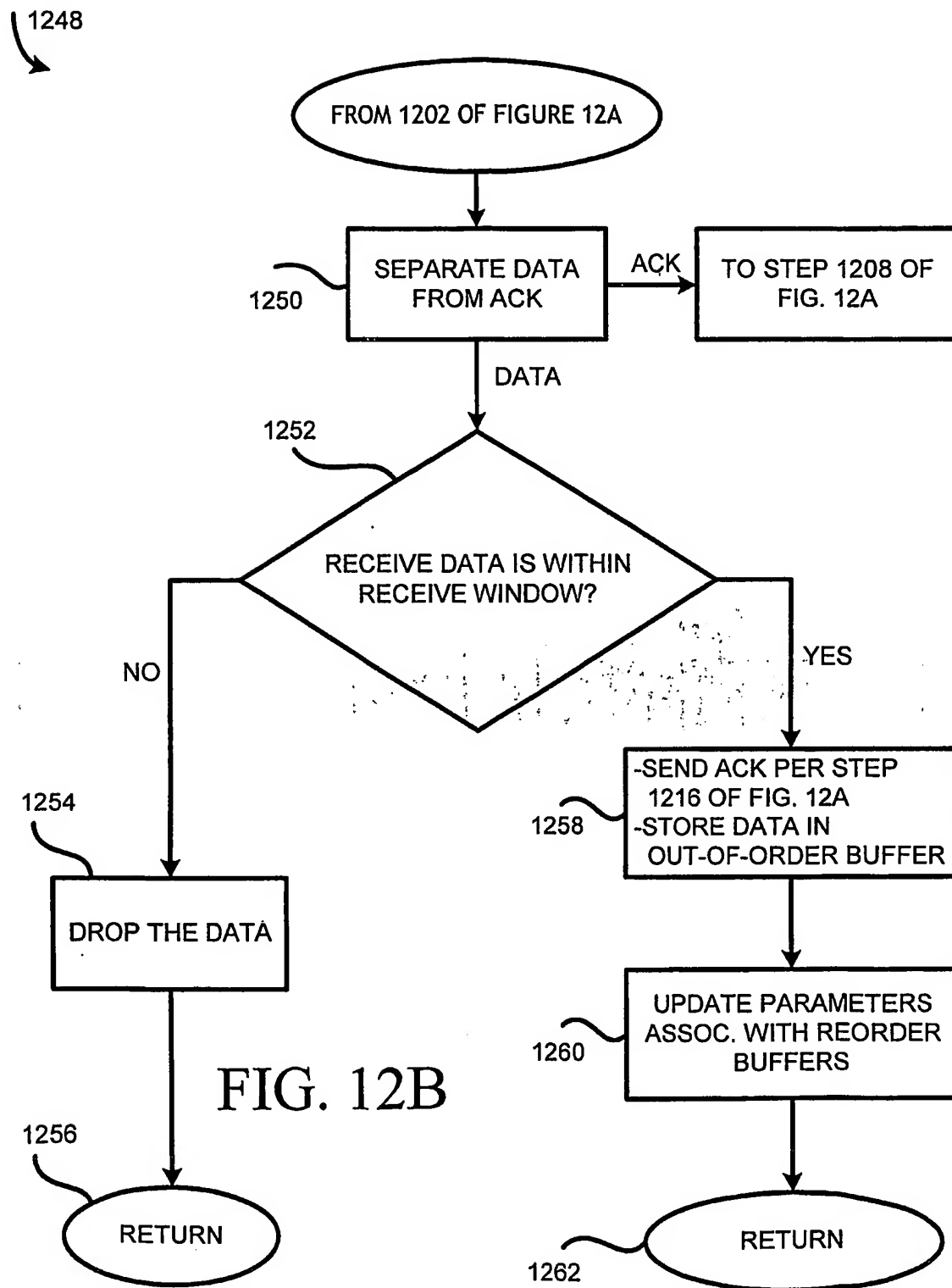


FIG. 11





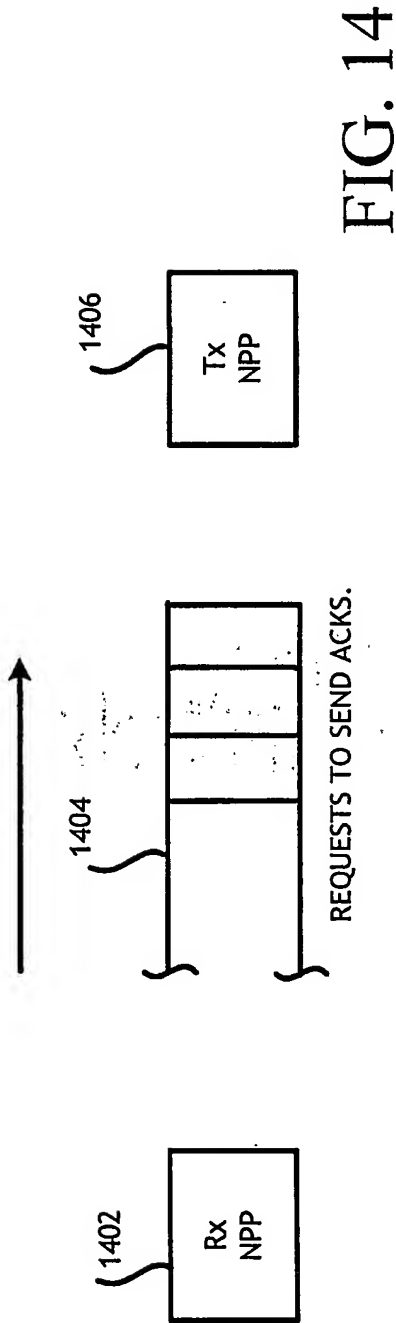
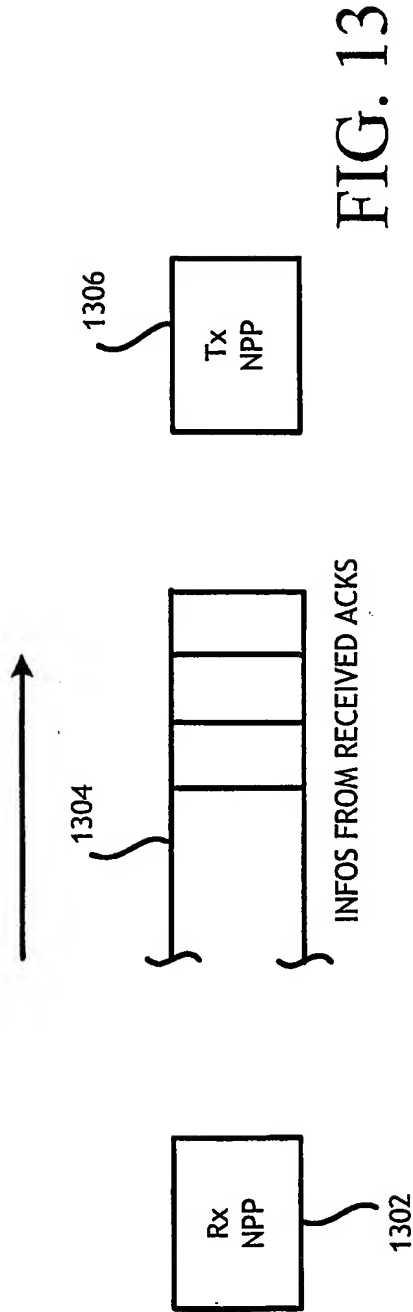


FIG. 15

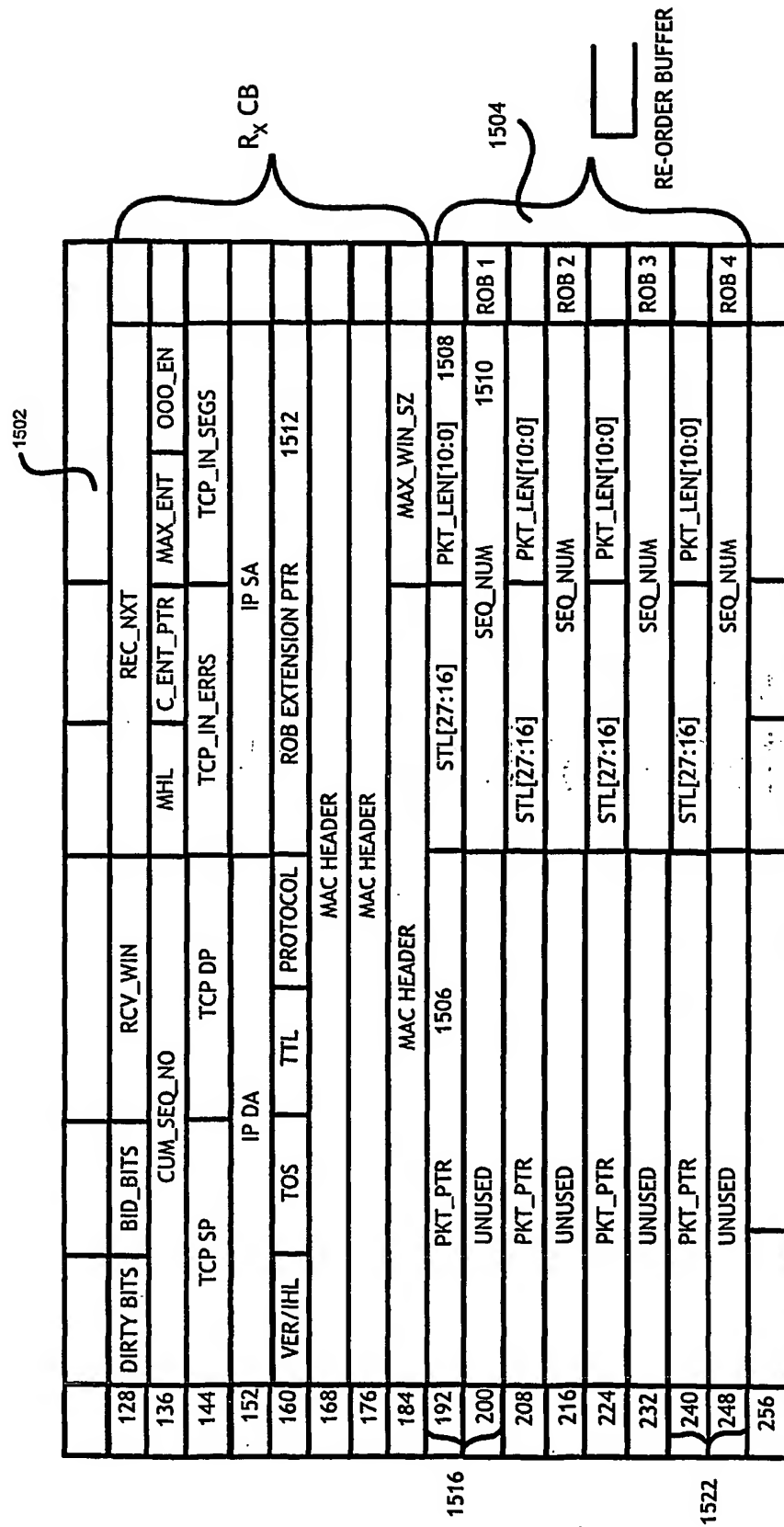


FIG. 16

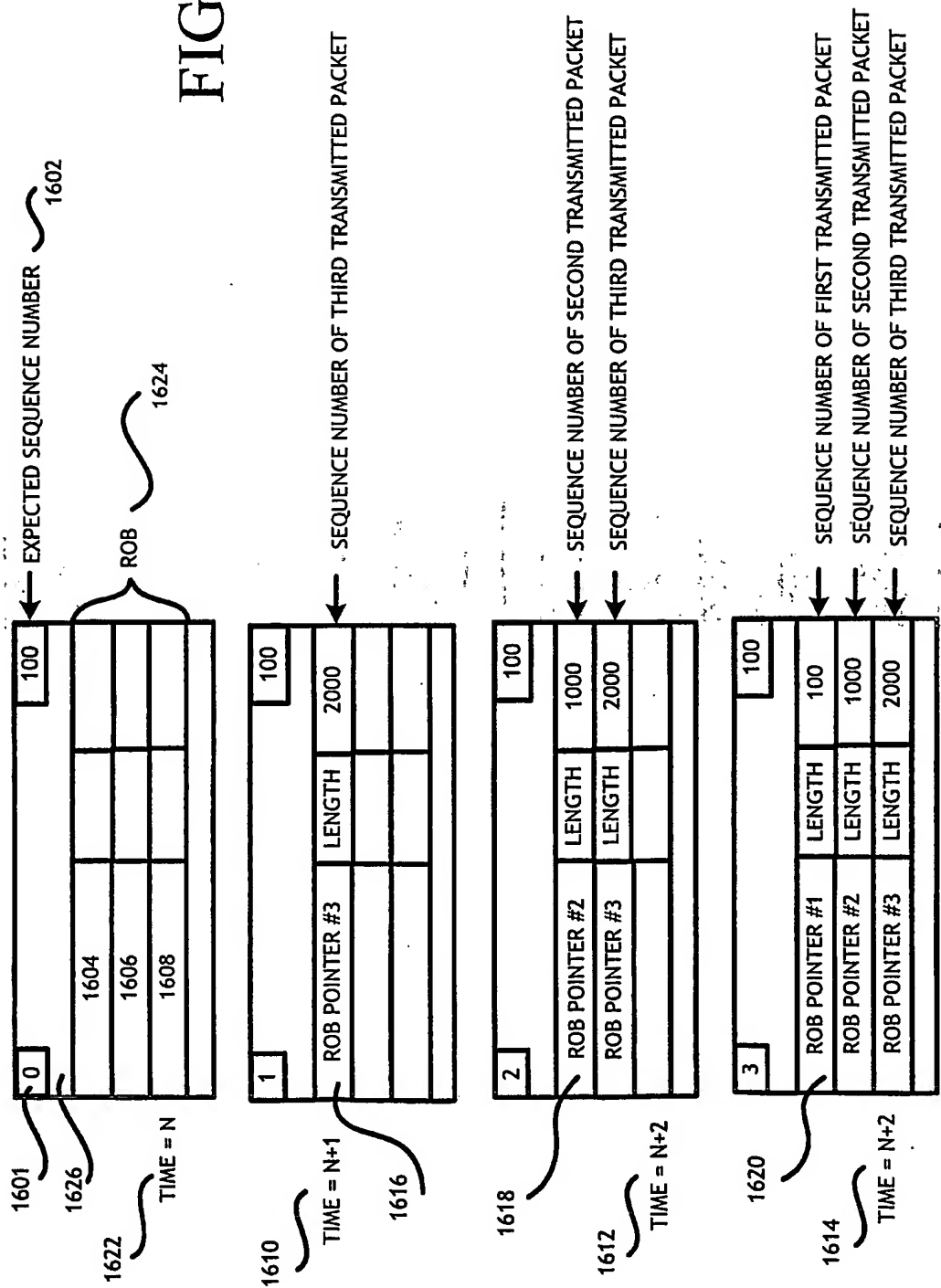


FIG. 17

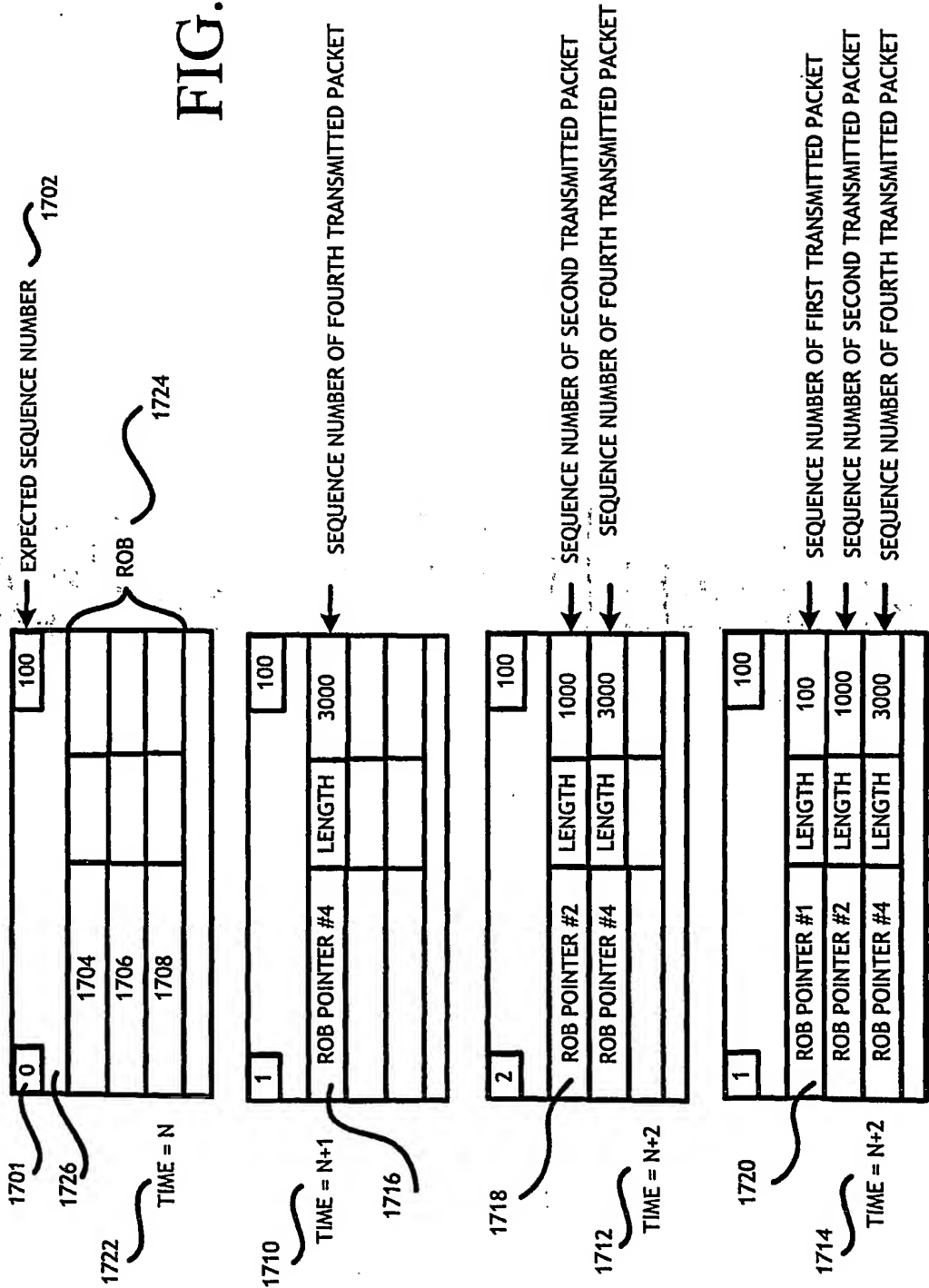


FIG. 18

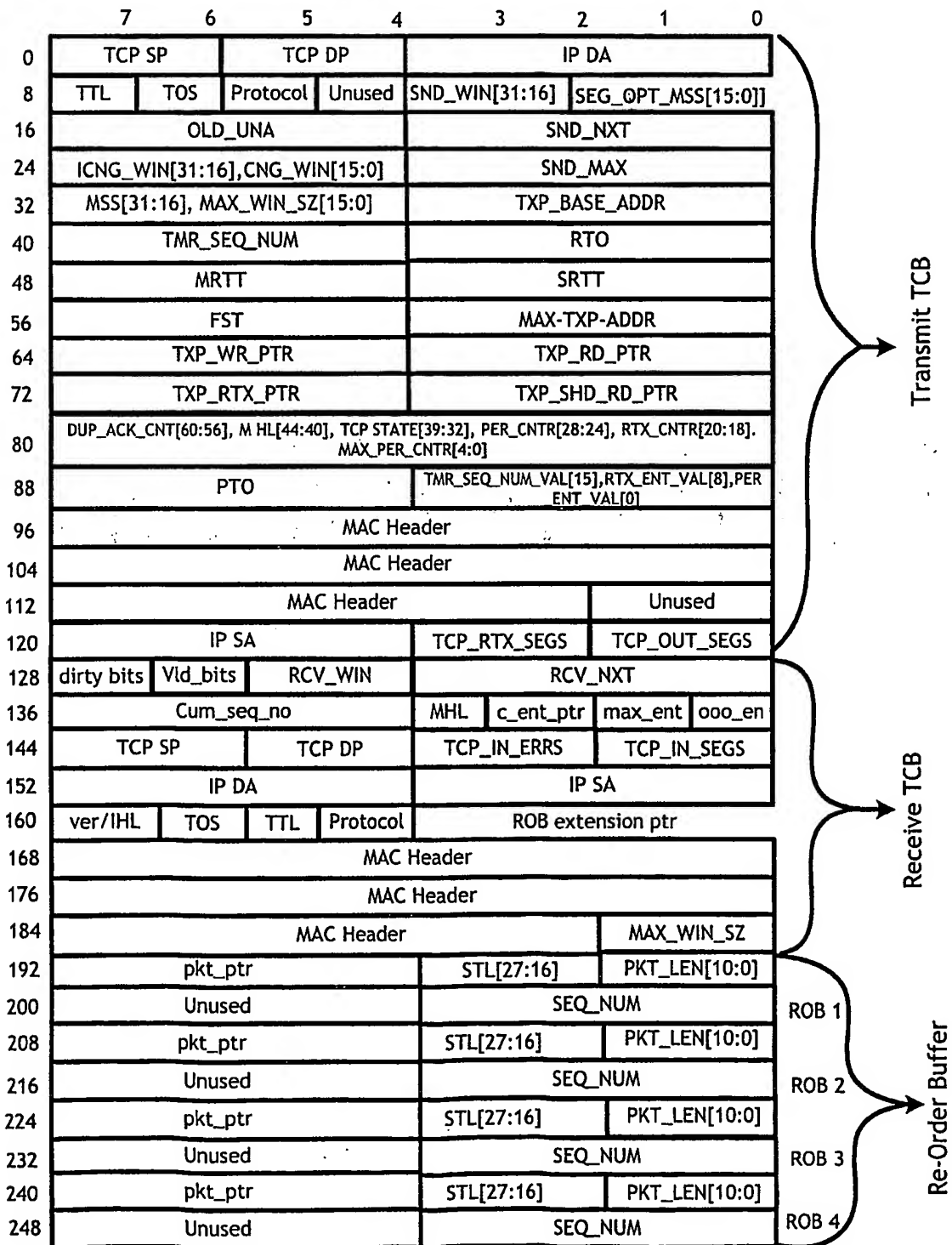
100			1808		
3	ROB POINTER #3	LENGTH	2000	ROB	
	ROB POINTER #4	LENGTH	3000		
	ROB POINTER #5	LENGTH	4000		

Source Port		Destination Port	
Sequence Number			
Acknowledgement Number			
HLEN	Reserved	Code Bits	Window
Checksum (15-0)		Urgent Pointer	
Options (if any)			Padding
Data			

FIG. A1
PRIOR ART

TCB LAYOUT IN MEMORY

FIG. A2A



FLOW CHART FOR REORDER BUFFER (ROB) UPDATE AND PROCESSING

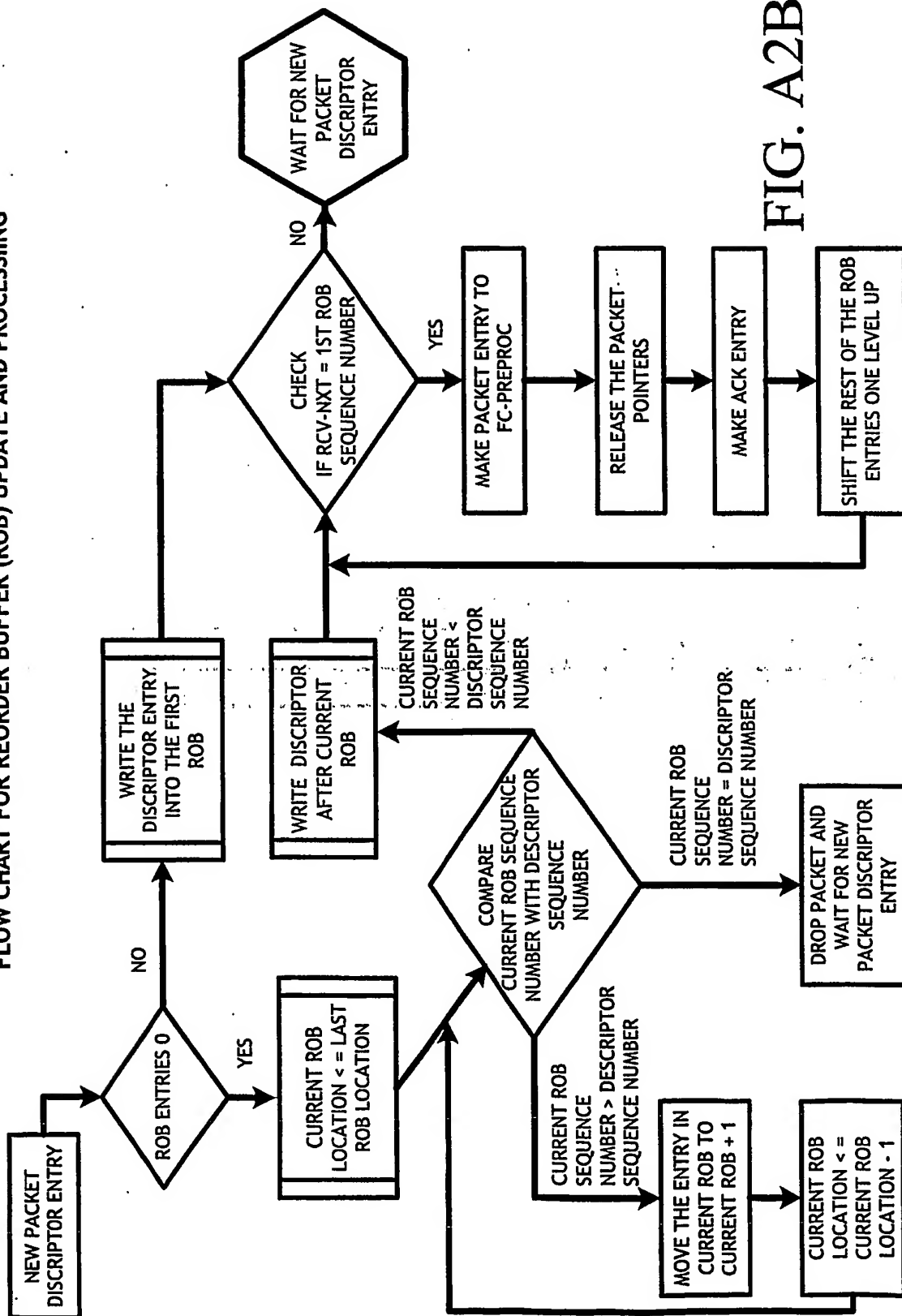


FIG. A2B

TCB Table***TCP Transmit Variables:***

TCB Field	Initial Value	Updated By	Neede d by SW TCB	Description
TCP_SP[15:0]	From SW		Yes	TCP Source Port.
TCP_DP[15:0]	From SW		Yes	TCP Destination Port.
IP_DA[31:0]	From SW		Yes	IP Destination Address.
TTL[7:0]	From SW		Yes	Time To Live. Used in IP Header.
TOS[7:0]	From SW		Yes	Type Of Service. Used in IP Header.
PROTOCOL[7:0]	From SW		Yes	Protocol Field. Used in IP Header.
SND_WIN[15:0]	From SW	Tcp_tx_proc, by tcp_rx_ack_proc when an ACK is received.	Yes	This is the window size offered by the sender.
SEG_OPT_MSS	From SW		No	Segment Optimal MSS.
OLD_UNA[31:0]	From SW	Tcp_tx_proc, by tcp_rx_ack_proc when ACK is received	Yes	Oldest Unacknowledged Sequence Number. Same as SND_UNA.
SND_NXT[31:0]	From SW	Tcp_tx_proc, by tcp_norm_cpu_proc, tcp_rx_ack_proc, when entry is put in port_tx_que	Yes	Sequence number to be sent next in the TCP header.
ICNG_WIN[15:0]	From SW		Yes	This is the value used for CNG_WIN when congestion control is used upon retransmission.
CNG_WIN[15:0]	From SW	Tcp_tx_proc, by tcp_rx_ack_proc for slow start on ACK receive, by tcp_rtx_proc when retransmission happens.	Yes	Current congestion window.
SND_MAX[31:0]	From SW	Tcp_tx_proc, by tcp_rx_ack_proc when ACK is received	Yes	Maximum Sequence Number allowed to be sent

FIG. A3A**SUBSTITUTE SHEET (RULE 26)**

TCB Field	Initial Value	Updated By	Neede d by SW TCB	Description
MSS[15:0]	From SW		Yes	Maximum Segment size. This value is aligned to 256-byte count.
MAX_WIN_SZ[15:0]	From SW		Yes	This is the window size which will be advertised as offered window when ACK is sent.
TXP_BASE_ADDR[31:0]	From SW		No	Base address for TX Pending queue. When the pointers roll over, they roll back to this value.
TMR_SEQ_NUM[31:0]	Don't Care	Tcp_tx_proc, when an entry is made in port_tx_que either by tcp_norm_cpu_tx_proc or by tcp_rx_ack_proc or after ACK validation if there are pending entries	No	Sequence number of the segment for which timer measurements are kicked off. The timers are updated when an ACK with ACK_NUM >= TMR_SEQ_NUM is received.
RTO[31:0]	Don't Care	Tcp_tx_proc, by tcp_rx_ack_proc after ACK validation	No	Retransmission Time Out value. It is updated every time an ACK with ACK_NUM >= TMR_SEQ_NUM is received.
MRTT[31:0]	Don't Care	Tcp_tx_proc, by tcp_rx_ack_proc after ACK validation	Yes	Measured Round Trip Time. It is updated every time an ACK with ACK_NUN >= TMR_SEQ_NUM is received.
SRTT[31:0]	From SW (In terms of number of 125 Mhz cycles.)	Tcp_tx_proc, by tcp_rx_ack_proc after ACK validation	Yes	Smoothened Round Trip Time. It is updated every time an ACK with ACK_NUN >= TMR_SEQ_NUM is received.

FIG. A3B

TCB Field	Initial Value	Updated By	Needed by SW TCB	Description
FST[31:0]	Don't Care	Tcp_tx_proc, when an entry is made in port_tx_queue either by tcp_norm_cpu_tx_proc or by tcp_rx_ack_proc	No	Frame Transmit Time. The current time at which the frame entry was made in the port_tx_queue. Obtained from a 32-bit counter.
MAX_TXP_ADDR[31:0]	From SW		No	Maximum address for TX Pending queue. When TXP pointers are greater than this address, they roll back to the starting address.
TXP_WR_PTR[31:0]	From SW (beginning of TXP queue)	Tcp_tx_proc, by tcp_norm_cpu_proc after an entry is serviced from pre_tx_queue.	No	Current Value of TX Pending queue write pointer.
TXP_RD_PTR[31:0]	From SW (beginning of TXP queue)	Tcp_tx_proc, by tcp_norm_cpu_proc or tcp_rx_ack_proc.	No	Current value of TX Pending queue read pointer.
TXP_RTX_PTR[31:0]	From SW (beginning of TXP queue)	Tcp_tx_proc, by tcp_rx_ack_proc	No	Current value of the TX Pending retransmit pointer. On a time out, entry pointed by RTX pointer is processed.
TXP_SHD_RD_PTR[31:0]	From SW (beginning of TXP queue)	Tcp_tx_proc, by tcp_norm_cpu_proc, tcp_rtx_proc	No	Current value of the TX Pending queue shadow read pointer. Used during retransmission of frames.
DUP_ACK_CNT[4:0]	From SW (zero)	Tcp_tx_proc, tcp_rx_ack_proc	Yes	Current value of duplicate ACK count.
MHL[4:0]	From SW			MAC Header Length.
TCP_TX_ST[3:0]	From SW	TBD	Yes	Current TCP transmit state.
TCP_RX_ST[3:0]	From SW	TBD	Yes	Current TCP receive state.
RTX_CNT[4:0]	From SW (zero)	Tcp_tx_proc, tcp_rtx_proc, tcp_rx_ack_proc	Yes	Current value of Retransmit counter.

FIG. A3C

SUBSTITUTE SHEET (RULE 26)

TCB Field	Initial Value	Updated By	Neede d by SW TCB	Description
PTO[31:0]	From SW (In terms of number of 125 Mhz cycles.)	Tcp_tx_proc, tcp_tx_pst_proc after sending a persist probe. If an ACK for the probe is not received and timeout occurs OR if an ACK with zero window size is received, then the persist probe is sent again and this time is doubled until MAX_PTO is reached. Then it maintains MAX_PTO value. It is reinitialized to INIT_PTO when an ACK with non-zero offered window size is received.	Yes	Persist Time Out value.
TMR_SEQ_NUM_VAL	From SW (zero)	Tcp_tx_proc. If reset, this bit is set by tcp_tx_norm_cpu_proc. Reset by tcp_rx_ack_proc when TX Pending queue is empty.	No	Timer Sequence Number Valid. When an ACK with ACK_NUM >= TME_SEQ_NUM is received, timer calculations are done. Also, if there are pending entries in the TXP, a new timeout value is written into RTX SRAM. When an ACK is received and there is nothing in the TX Pending queue, then this bit is reset.
RTX_ENT_VAL	From SW (zero)	Tcp_tx_proc, If this bit is reset, then it is set by the process making an entry in the port_tx_que.	No	Retransmit entry valid. When set, it means that the RTX timer is running.
PST_ENT_VAL	From SW (zero)	Tcp_tx_proc, by tcp_rx_ack_proc. Set when an ACK with zero offered window is received. Reset when an ACK with non-zero offered window is received.	No	Persist Entry Made.

FIG. A3D

TCB Field	Initial Value	Updated By	Needed by SW TCB	Description
MAC[191:0]	From SW		No	MAC Header. Maximum possible header length is 22-bytes. The actual length is given by MHL. This field appears twice so it can be accessed by the RX engine for outgoing ack packets.
TCP_RTX_SEGS[15:0]	From SW	Tcp_tx_proc, tcp_rtx_proc when a segment is retransmitted. Need to calculate the segmented packets.	Yes	Total number of segments retransmitted. (takes into account segmentation).
TCP_OUT_SEGS[15:0]	From SW	Tcp_tx_proc, by tcp_nom_cpu_tx_proc or tcp_rx_ack_proc or by tcp_rtx_proc or by tcp_tx_pst_proc, any time an entry is made in the port_tx_queue. Need to calculate the total segmented packets.	Yes	Total number of TCP segments sent out on this session including the segmented packets.

Receive Variables:

TCB Field	Initial Value	Updated By	Needed By SW TCB	Description
RCV_WIN[15:0]	From SW	Tcp_rx_proc	Yes	Current Receive window. When tcp_rx_proc sends an ACK, this value is used in the offered window field in the TCP Header.
RCV_NXT[31:0]	From SW	Tcp_rx_proc	Yes	The next expected sequence number by tcp_rx_proc.
cum_seq_no[31:0]	From SW (= RCV_NXT)	Tcp_rx_proc	No	Accumulated Segment Length. Used in frame reassembly.
vld_bits[7:0]]	From SW	Tcp_rx_proc	No	Valid 64-byte segments in the TCB, each of the 8

FIG. A3E
SUBSTITUTE SHEET (RULE 26)

TCB Field	Initial Value	Updated By	Need d By SW TCB	Description
	(0x04)			64-byte segments has a corresponding bit.
dirty_bits[7:0]	From SW (0x00)	Tcp_rx_proc	No	For each of the 8 64-byte TCB/ROB segments a corresponding dirty bit indicates it has been updated and needs to be flushed back to memory..
OOO_EN	From SW	SW	No	When enabled, out of order packets are processed, other wise only sequential packets. are processed.
max_ent[3:0]	From SW	Tcp_rx_proc	No	maximum no. of entries in the ROB, can be upto 20
c_ent_ptr[3:0]	From SW (0x0)	Tcp_rx_proc	No	current entry pointer. Points to next entry to be filled.
TCP_IN_SEGS[15:0]	From SW (0x0)	Tcp_rx_proc	No	current no. of TCP segments received so far.
TCP_IN_ERRS[15:0]	From SW (0x0)	Tcp_rx_proc	No	current no. of TCP segments with errors received so far.
MAC Header	From SW	SW	No	MAC header for use by outgoing ack packet
MHL	From SW	SW	No	MAC Header Length for use by the outgoing ack packet
TCP_SP[15:0]	From SW		Yes	TCP Source Port.
TCP_DP[15:0]	From SW		Yes	TCP Destination Port.
MAX_WIN_SZ[15:0]	From SW		Yes	This is the window size which will be advertised as offered window when ACK is sent.
TTL[7:0]	From SW		Yes	Time To Live. Used in IP Header.
TOS[7:0]	From SW		Yes	Type Of Service. Used in IP Header.
PROTOCOL[7:0]	From SW		Yes	Protocol Field. Used in IP Header.

FIG. A3F

SUBSTITUTE SHEET (RULE 26)

TCB Field	Initial Value	Updated By	Neede d By SW TCB	Description
IP_SA[31:0]	From SW		Yes	IP Source Address.
IP_DA[31:0]	From SW		Yes	IP Destination Address.

Re-Order Buffer:

TCB Field	Initial Value	Updated By	Neede d By SW TCB	Description
FRM_LEN[15:0]	Don't Care	Tcp_rx_proc when out of order packets are received.	No	Frame length for the out of order segment.
PKT_LEN[15:0]	Don't Care	Tcp_rx_proc when out of order packets are received.	No	Length of the current segment.
SEQ_NUM[31:0]	Don't Care	Tcp_rx_proc when out of order packets are received.	No	Sequence Number of out of order segment.
SAF[31:0]	Don't Care	Tcp_rx_proc when out of order packets are received	No	Start Address where the out of order segment is stored in SDRAM1.

FIG. A3G

FIG. A4

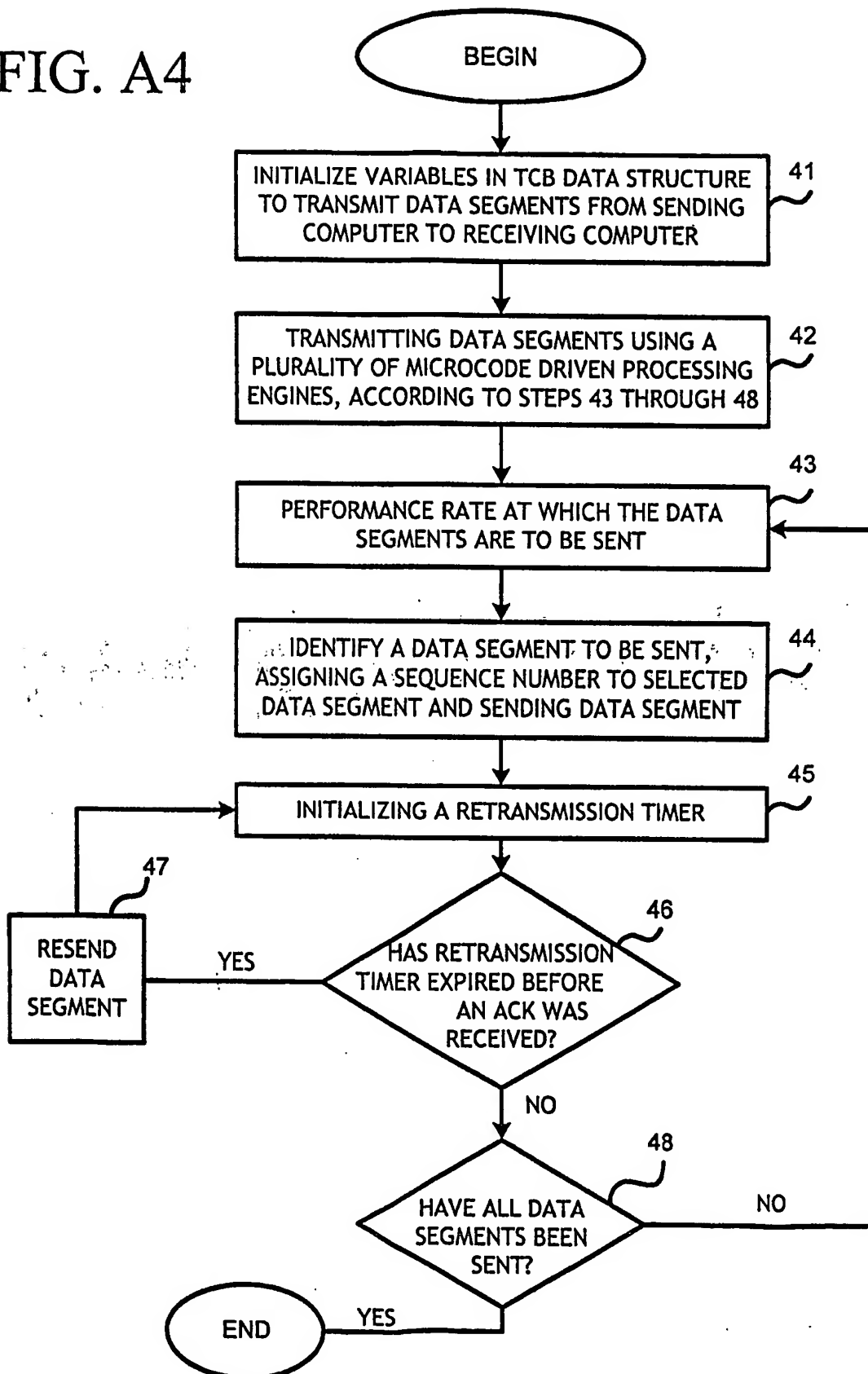
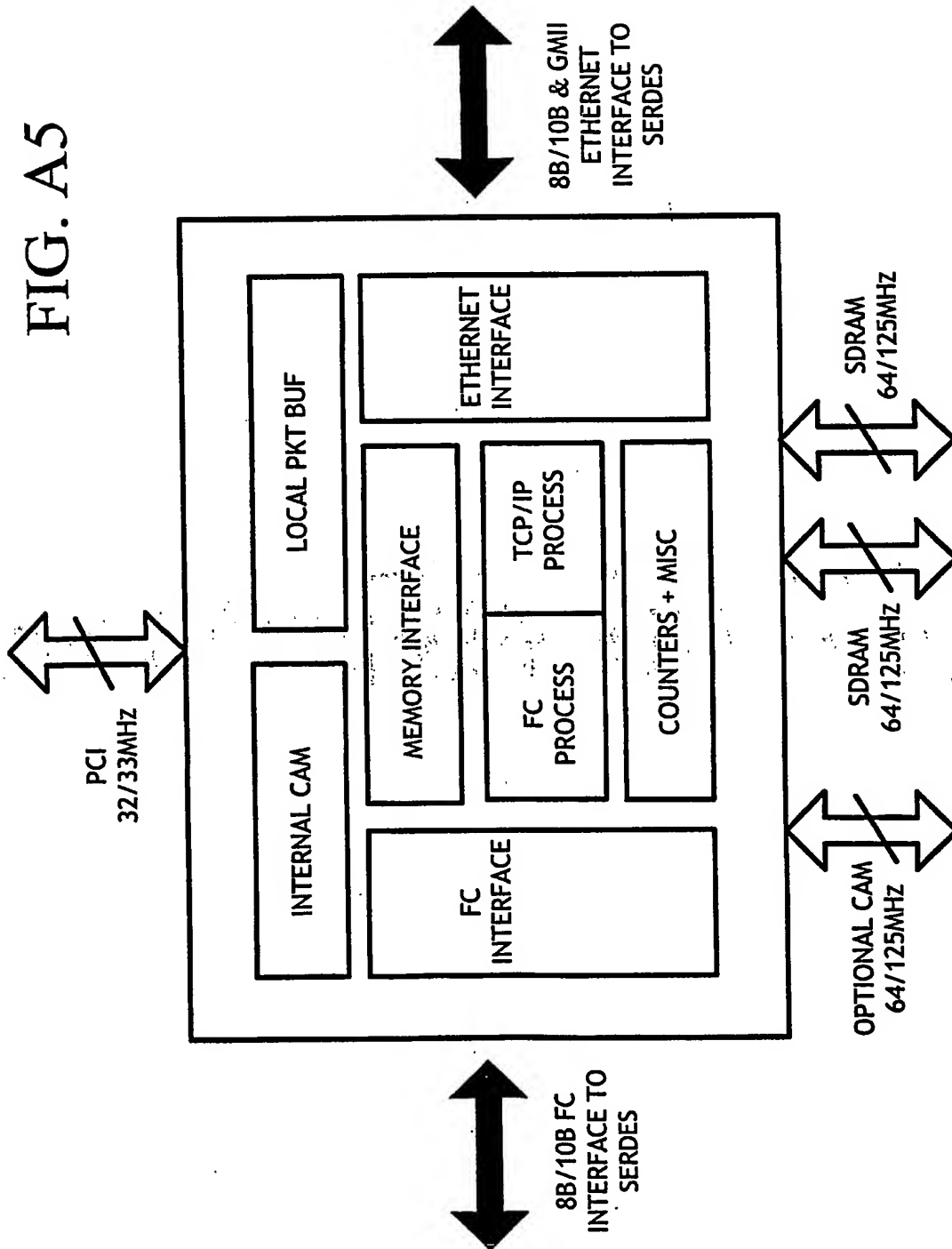


FIG. A5



FC to GE Processing

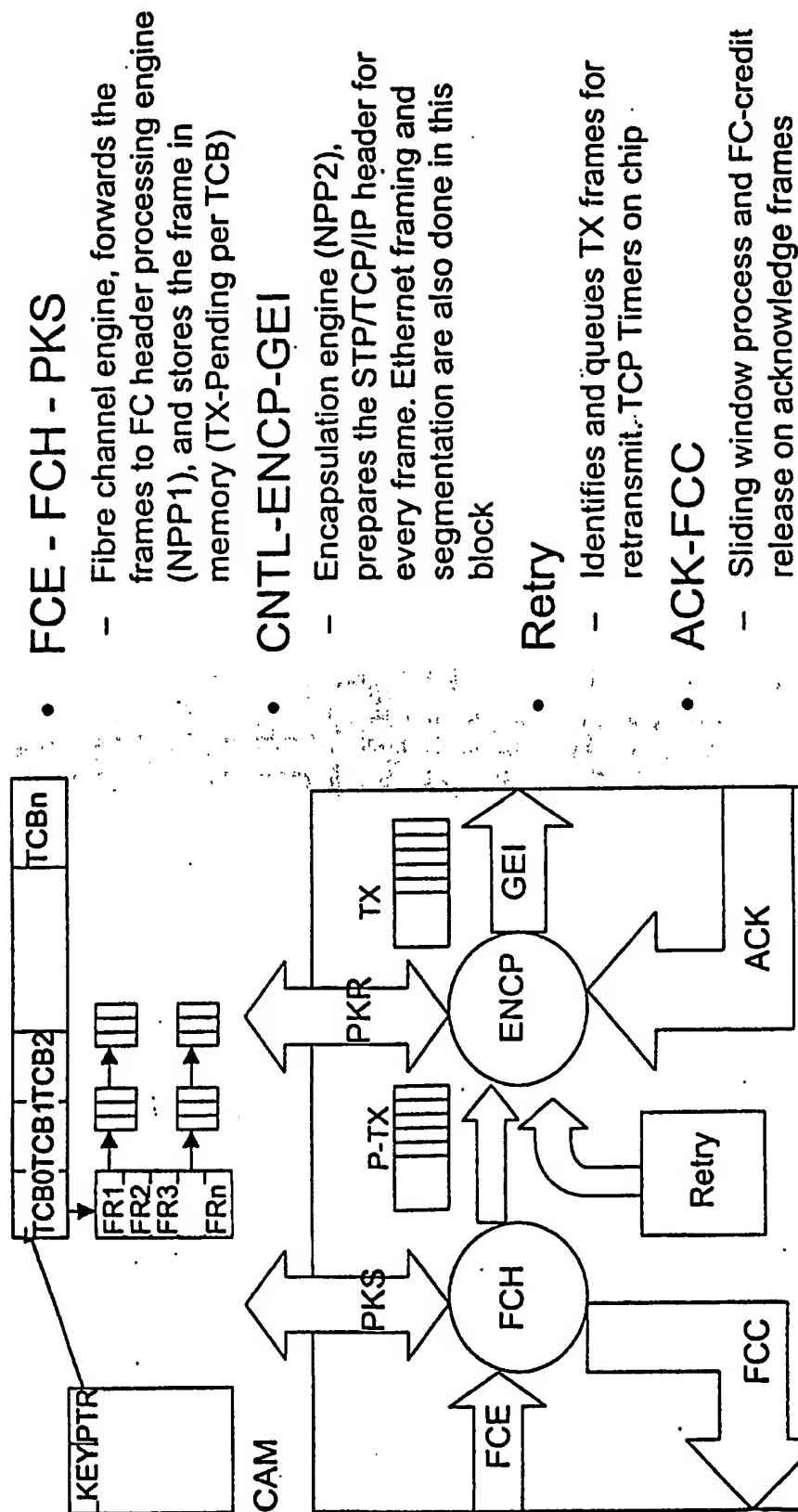


FIG. A6

GE to FC Processing

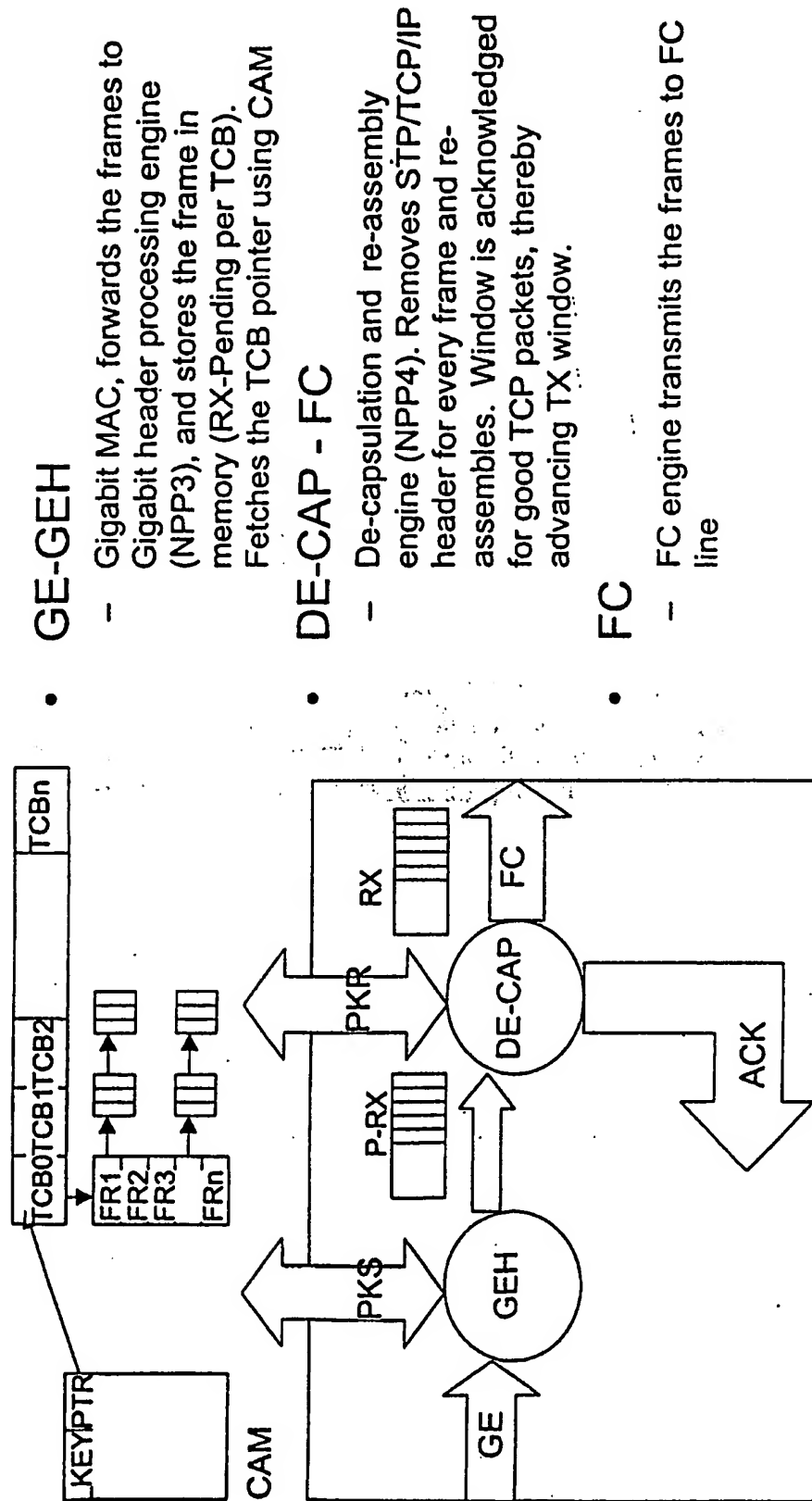


FIG. A7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/27707

A. CLASSIFICATION OF SUBJECT MATTER		
IPC(7) : G06F 13/00 US CL : 709/200, 230, 227, 228 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols) U.S. : 709/200, 230, 227, 228		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) WEST USPAT		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,600,793 A (NORD) 04 February 1997, Abstract, col. 2, lines 30-50	1-10
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 02 December 2002 (02.12.2002)		Date of mailing of the international search report 18 DEC 2002
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230		Authorized officer Meng A An <i>Juan R. Martinez</i> Telephone No. 703 305-9700